Idaho Ground Water Quality Rule

Statistical Guidance

July 2007

Idaho Ground Water Quality Rule

Statistical Guidance

July 2007

Prepared by Dr. Xin Dai, Idaho Department of Environmental Quality
Technical Review Provided by Dr. John Welhan, Idaho Geological Survey
Edited by Edward Hagan, Idaho Department of Environmental Quality

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Acronym/Symbol Definition List

α	False rejection (or false positive) decision error
ACL	Alternative concentration limit
b_1	Slope of the linear regression line
b_0	Intercept of the linear regression line
COC	Constituents of concern
CV	Coefficient of variation
γ	Skewness
IDEQ	Idaho Department of Environmental Quality
EPA	U.S. Environmental Protection Agency
F	Variance ratio from the table of the F-distribution
H_0	Null hypothesis
H_A	Alternative hypothesis
IQR	Interquartile Range
K	Kruskal-Wallis (K-W) test statistic
k	Number of seasons (typically 4 for the K-W seasonality test)
K	Multiplier used for setting UTLs or PLs
k	The number of future comparisons
m	The number of years for which data were collected
MSE	Mean square error
N	The sample size or total number of measurements (= n x m)
n	The number of measurements per year (quarterly = 4)
ppm	Parts per million
r^2	Coefficient of determination
$\overline{R}_j$	Average group rank for Kruskal-Wallis test
s	The standard deviation of a sample data set
S	The Mann-Kendall test statistic
s^2	The variance of a sample data set
SSE	Sum of squares due to error
$s_{\bar{x}}$	Standard error
TDS	Total dissolved solids
UPL	Upper prediction limit
UTL	Upper tolerance limit
W	Shapiro-Wilk test statistic
W	Levene test statistic
WLAP	Wastewater land application permit
x_i, y_i	Constituent concentration for the i^{th} ground water sample
$\overline{x}$ or $\overline{X}_N$	The mean (or average) of a sample data set
$\overline{x}_k$	The mean for all values from the same month but different years
x_{jk}	An alternative way of denoting a chemical measurement, where k = 1, 2, …, m denotes the year, and j = 1, 2, …, n denotes the sampling period (season) within the year. The subscript for x_{jk} is related to the subscript for x_i in the following manner: i = (k-1)n + j.
$\chi^2_{1-\alpha,(k-1)}$	The 1-α quantile of a chi-square distribution with k-1 degrees of freedom

I. Introduction

This guidance describes a process for the Idaho Department of Environmental Quality (DEQ) to use when determining if ground water quality is degraded. The term "degradation" is defined in the Idaho Ground Water Quality Rule (IDAPA 58.01.11) "as the lowering of ground water quality as measured in a statistically significant and reproducible manner." This guidance provides a process and statistical tools which can be used to determine statistically significant degradation.

The two principal goals of the guidance are:

- describe a statistically based process for establishing background ground water quality; and
- identify methods and criteria for identifying when ground water quality degradation is statistically significant.

An understanding of these two concepts is fundamental to addressing ground water quality issues. Knowledge of the background ground water quality is necessary before ground water quality degradation can be identified. Once background ground water quality is established, ground water quality degradation can be determined.

To achieve the two principal goals the guidance is structured around the following four objectives:

- provide a standardized framework or process to objectively evaluate ground water quality data;
- utilize a decision tree showing required elements;
- provide flexibility to address site-specific conditions; and
- suggest certain statistical tools but allow for alternatives.

The determination of what constitutes degraded ground water is essential for implementation of DEQ programs that rely on the Idaho Ground Water Quality Rule (Rule) to protect the health of Idahoans and the environment. DEQ and the regulated community can utilize the methods contained in this document to estimate background ground water quality conditions and identify degradation. This guidance document is intended to help interpret and apply the Idaho Ground Water Quality Rule at sites not addressed with existing state and federal program guidance. It may also complement existing guidance by addressing situations not covered with other guidance. This document does not impose legally binding requirements on DEQ or the regulated community. The document identifies an approach for defining ground water degradation, but DEQ retains the discretion to adopt approaches on a case-by-case basis that differ from this information. Interested parties are free to raise questions about the appropriateness of the application of the information in this document to a particular situation, and DEQ will consider whether or not the technical approaches are appropriate in that situation.

II. Authorities and Definitions

The legislation and rules addressing ground water quality issues in Idaho include the Idaho Ground Water Quality Protection Act of 1989 (Act) (§ 39-120 to § 39-127, Idaho Code) and the Idaho Ground Water Quality Rule (Rule) (IDAPA 58.01.11). The Act created the Ground Water Quality Council and directed the Council to develop the Ground Water Quality Plan. The Plan provides the overall direction and policies of the state with respect to ground water quality concerns. The Rule implements a portion of the Plan.

Background water quality can be established using samples collected from monitoring wells that sample the ambient ground water quality in the same aquifer that is likely to be impacted by development. The Rule identifies two types of background ground water quality: natural and site.

Natural background level is defined by the Rule "as the level of any constituent in the ground water within a specified area as determined by representative measurements of the ground water quality unaffected by human activities." In areas where the natural background level of a constituent exceeds the standard, the natural background level shall be used as the standard.

Site background level is defined as the ground water quality at the hydraulically upgradient site boundary. In areas where the ground water quality is unaffected by human activities, the site background level is equivalent to natural background.

III. Statistical Characterization of Background Water Quality

Before any data evaluation begins it is useful to have a clear understanding of the issues that need to be addressed, including the constituents of concern. Once the main issues are defined, the data can be collected, and then reviewed within the appropriate context. Existing data must be compiled and evaluated to determine if the information is sufficient to adequately characterize the ground water quality. In most cases, the goal of the statistical analysis will be to characterize background ground water quality in a valid manner such that decisions regarding ground water quality degradation are defensible.

The guidance provides flexibility by allowing options to determine background ground water quality, depending on the adequacy of the data for statistical analysis. If sufficient data are available to statistically characterize background water quality, then appropriate statistical methods may be employed. Suggested methods to determine if the available ground water quality data are adequate to conduct valid statistical analyses are described in the Appendices.

If data are insufficient to conduct valid statistical analyses, then a sampling plan to collect adequate data may be developed or background ground water quality may be estimated using an alternative concentration limit (ACL) in accordance with a DEQ prescribed method. The ACL is designed to be protective of ground water quality by using the lowest value provided from three options as described in Appendix A. If a sampling plan is implemented, the ACL will be used for decision-making purposes until adequate data

are collected to support valid statistical analyses. However, an ACL also may be selected even when appropriate data are available for valid statistical analyses if the interested party does not want to conduct a statistical analysis and DEQ concurs with the decision.

III.1 Elements of an Analysis

The elements for characterizing background water quality are site-specific and dependent on the complexity of the area. The steps to be completed for each site include:

- State the objectives of the analysis;
- Delineate the study area and hydrogeologic features relevant to monitoring;
- Identify constituents of concern and provide rationale for considering them
- Evaluate and define data adequacy in the context of the analysis objectives;
- Identify appropriate statistical tools to address the issues;
- If the data are inadequate for the analysis, determine an appropriate temporal scale for the data collection program and provide a rationale for why it is appropriate; and
- If selected data are used in (or excluded from) the analysis, provide a rationale.

The general process for defining background ground water quality is illustrated in the flow diagram in Figure 1.

The elements must be addressed within the context of the hydrogeologic framework. Individual aquifers must be defined at the appropriate scale. For each aquifer the ground water flow direction and ground water gradient should be described and uncertainties in both should be estimated. Data on the ground water chemistry of each aquifer should be compiled and ground water quality trends should be identified, if data are sufficient. The sampling locations and sampling frequency should be evaluated to ascertain if results can be used to represent the ground water quality within the area of concern.

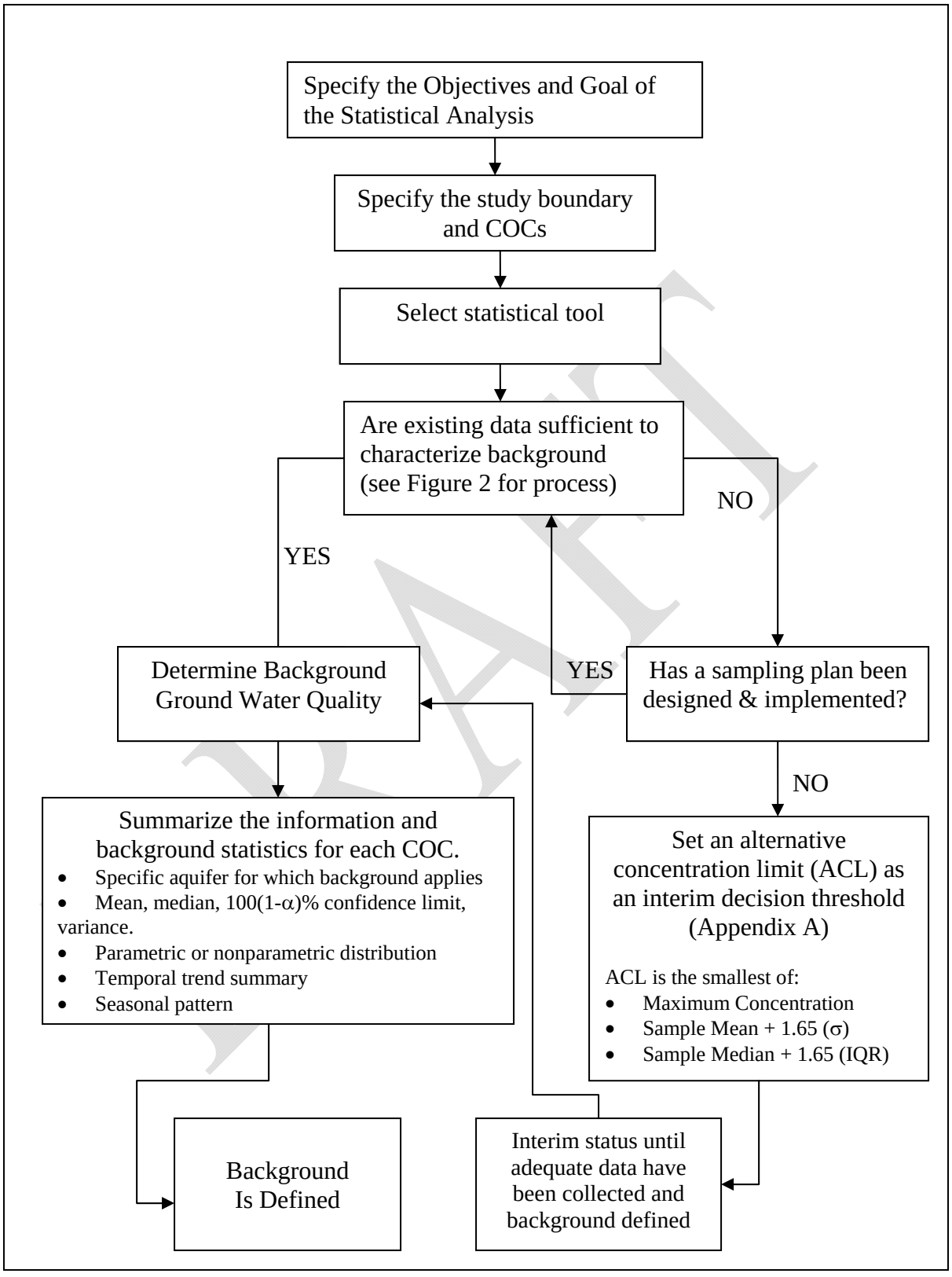


Figure 1. Process for determining background ground water quality

III.2 Evaluation of Hydrogeologic Data

Ambient ground water quality typically varies spatially (between wells) and temporally (over time) at individual wells due to natural conditions. Anthropogenic activities may contribute to the variability observed in water quality data. Though a minimum of one upgradient well is needed to determine background water quality, spatial variability cannot be accurately assessed without at least three up-gradient wells (Fisher and Potter, 1989; Cressie, 1993; Gibbons, 1994). For sites that have not been previously affected by human activities, both upgradient and downgradient wells can be used to determine background water quality and spatial variability.

The number of downgradient monitor wells should be sufficient to provide an accurate representation of the ground water quality to ensure that appropriate management decisions are made. Some downgradient wells need to be placed at the point(s) of compliance for the property. It is important to understand that since the statistical tests are based on data collected from monitoring wells that are in fixed locations, the statistical conclusions refer only to water in the monitoring wells and not in the aquifer in general (EPA, 1988; EPA, 1992b).

Since hydraulic anisotropy causes ground water to flow in a direction that is not always perpendicular to measured water table gradients (Domenico and Schwartz, 1990; Fetter, 1993), a line of downgradient wells at the discharge boundary of the site would provide the best chance for detecting the highest value(s) generated at the site. Idaho's *Guide to Ground Water Sampling and Monitoring* (Ogden, 1987*)* suggests a downgradient spacing of 150 feet between wells at hazardous waste disposal sites, with greater spacing at non-hazardous waste disposal sites. The monitoring plan should have sufficient wells to characterize the entire site in question.

The hydrogeologic characteristics of the site will play a primary role in determining the number and location of the ground water monitoring wells. The depth to water and flow direction, net recharge, aquifer material, soil characteristics, topography, thickness and lithology of the vadose zone, and hydraulic conductivity or permeability of the aquifer are all important in determining pollution potential of an aquifer and the necessary spacing and depth of monitoring wells (Ogden, 1987). The geology of a site should be characterized through the interpretation of well logs, geologic maps, and cross sections. Structural features such as faults, fractures, fissures, impermeable boundaries, or other features that can influence flow direction, should be delineated. Additional hydrogeologic information, if available, should be summarized where it is relevant to the adequacy of monitoring data, including but not limited to ground water flow velocity, transmissivity, storage coefficient, porosity, and dispersivity.

III.3 Defining Constituents of Concern

A constituent of concern (COC) is a chemical that is disturbed, generated, used or disposed at the site in sufficient quantity to pose a risk to beneficial uses of ground water or interconnected surface water. This includes degradation products or chemicals released during chemical reactions in the environment. COCs must be defined for each

site and will be dependent on the site operations. When deciding which chemical may be a COC, the following should be carefully considered:

- The industrial/commercial processes resulting in the generation of the chemical(s) that are permitted to be handled, stored, or land applied on the site;
- The physical and chemical properties of the chemical(s);
- The methods of sample collection, handling, and transportation are appropriate for the COC;
- The laboratory analysis procedures used to measure chemical concentration are appropriate for the COC; and
- The complexity and sensitivity of the hydrogeologic environment.

III.4 Adequate Sample Size

This section specifically addresses quantifiable measurements above the detection limit not affected by censoring. Procedures for dealing with censored data are discussed in section III.5. The quality and quantity of available monitoring data are two of the most important factors in determining background water quality for a COC. Individual ground water samples are only representative of ground water quality at a particular time in a particular location. Ground water quality often varies seasonally or changes with time and/or location, so a single ground water sample may not be representative of ground water conditions throughout the site or over a period of time. The greater the number of independent samples collected over time, the more representative the characterization of the ground water quality. Larger sample populations also increase the statistical confidence in the evaluation of ground water quality. Statistical testing depends upon collection of adequate data. Statistical tests are based on using estimates of the true mean and true variance of a population. For example, the estimate of the true mean is the average of the data points collected. The estimate of the true standard deviation is the standard deviation of the data points collected.

The number of independent samples is dependent on the site-specific conditions, which in turn controls the data variability. Under ideal circumstances, the U.S. Environmental Protection Agency (EPA 1992a) asserts that there must be 8 to 10 independent samples before one can generate a passable estimate of the population standard deviation for populations having normal or lognormal (parametric) distributions. DEQ recommends collecting 12 independent samples for most statistical analysis methods discussed in this guidance document. In contrast, estimating a tolerance interval for populations that are non-parametrically distributed, requires a minimum of 59 independent data points for 95% coverage (where 95% of the data fall within the interval), at 95% confidence (Conover, 1999; EPA, 1992a; Gibbons, 1994, p.93).

In situations where a seasonal trend is present within the data set, the Seasonal Kendall Test requires a minimum of three years (36 data points) of monthly data (Gilbert, 1987, p.225). Harris et al. (1987) state that one is unlikely to be able to quantify serial correlation (independence) in quarterly ground water data without at least 10 years of quarterly data, or 40 data points. When quarterly data are sparse, the Kruskall-Wallis

Test can be used as long as there are at least three years of quarterly data taken in the same months (a minimum of 12 independent data points).

For example, DEQ recommends facilities collect a minimum of 12 independent quarterly samples for the determination of background water quality in each monitoring well for each COC. This enables the monitoring program to capture the impact of seasonal fluctuations which typically occur at a site. Collecting data above these minimum requirements will be useful and may be necessary in certain instances to better characterize the background water quality.

III.5 Data Below Detection Limits

A component of many data sets is non-detect values. These data are referred to as censored data throughout the remainder of this guidance. For most nonparametric methods censored data is not an issue, but in parametric analyses, the effect of censored data is very dependent on the statistical form of the distribution of the data. The procedure to evaluate censored data provided below in Figure 2 therefore applies primarily to parametrically distributed data. Additional description of methods for handling censored data are provided in Gibbons (1994) and Helsel (1990, 2005).

The first step when evaluating censored data is to distinguish between detection only applications such as identifying the first arrival of a constituent and ground water quality characterization (e.g. defining background or determining compliance). If the data are used to determine whether a constituent is present, the results should be handled on a case-by-case basis independent of the process outlined in Fig. 2. If the censored data will be used to estimate summary statistics, the procedure outlined in Figure 2 is applicable.

The percentage of censored data within the entire data set determines the preferred method of describing the censored data. However, before any analyses are performed DEQ recommends that the data set consist of at least 12 uncensored independent measurements to ensure the results of the analysis are representative of the ground water quality.

Depending on the proportion of censored data, the recommended procedure utilizes one of the three basic methods for estimating summary statistics for censored data summarized by Helsel (1990). These methods include:

- simple substitution;
- distributional methods; and
- robust methods.

If the percentage of censored measurements is relatively small, then simple substitution methods have been shown to be satisfactory for estimating parameters (Helsel, 1990). If the percentage of censored data represents less than 20% of the data set then ½ the censoring limit is a valid estimate of the censored data values. The censoring limit can vary depending on the project and the laboratory analytical methods used. It may include the practical quantitation limit, the method reporting limit, or the method detection limit.

If the percentage of censored measurements is greater than 20%, but less than 40%, and the data are deemed to be normally or lognormally distributed then the statistical parameters of the distribution may be inferred using distributional methods described in Helsel (1990), such as the maximum likelihood estimator (MLE).

If more than 40% of the data set is comprised of censored measurements and the data are parametrically distributed, then multiple methods, including robust methods (Helsel, 1990) should be evaluated to estimate the distribution's parameters, including a sensitivity analysis of the results to decide on the best method.

The Department may approve statistical methods for handling censored data other than those outlined in Figure 2.

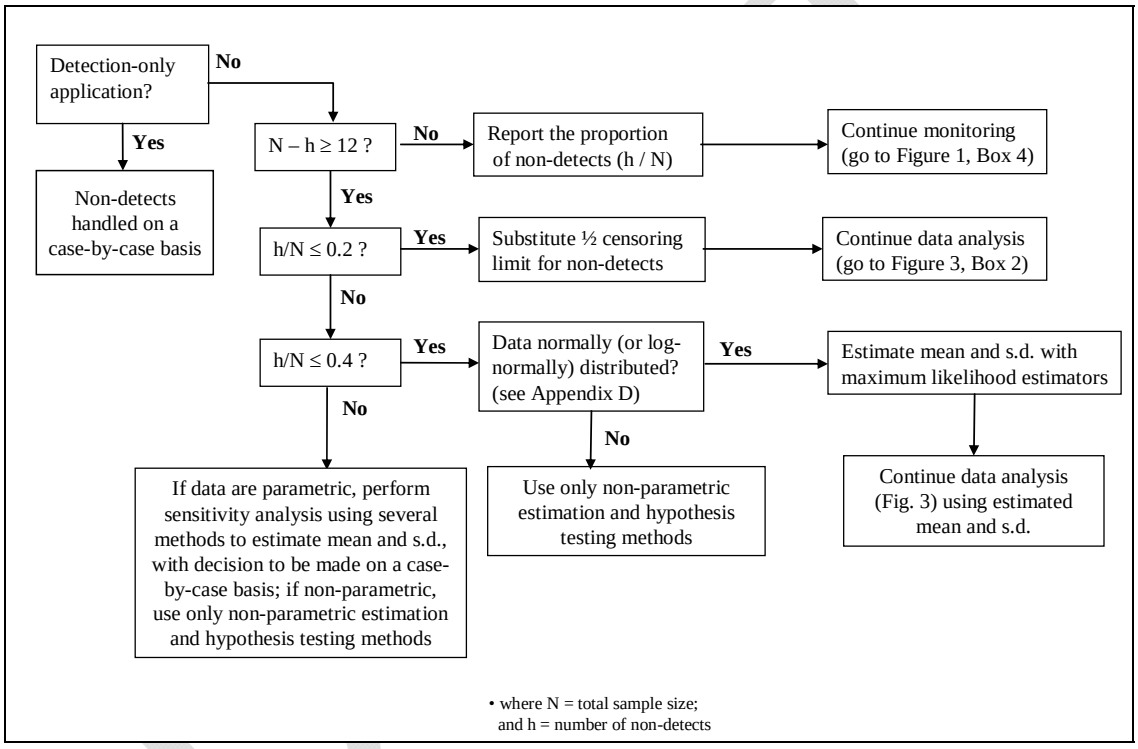


Figure 2. Suggested Procedural Guidance for Handling Censored Data

III.6 Evaluation of Background Water Quality Data

The procedure for evaluating data to determine its suitability for statistical analysis and the information and analysis required to substantiate a statistical characterization of background water quality is outlined in Figure 3. The steps include: data compilation; exploratory analysis and descriptive statistics; evaluation of data independence; analysis of frequency distribution and parametric behavior; seasonal and secular trend analysis; justification for data pooling if used; and an assessment of the adequacy of the available data to support a statistical characterization of background water quality (see Section V for more details on these terms). It is necessary to accurately characterize background water quality based on a sufficient number of samples to determine average

concentrations and variability at the site. Background water quality should be calculated using the most current data available.

Finally, if the available data are deemed adequate to justify such an analysis, summarize the results of the statistical characterization.

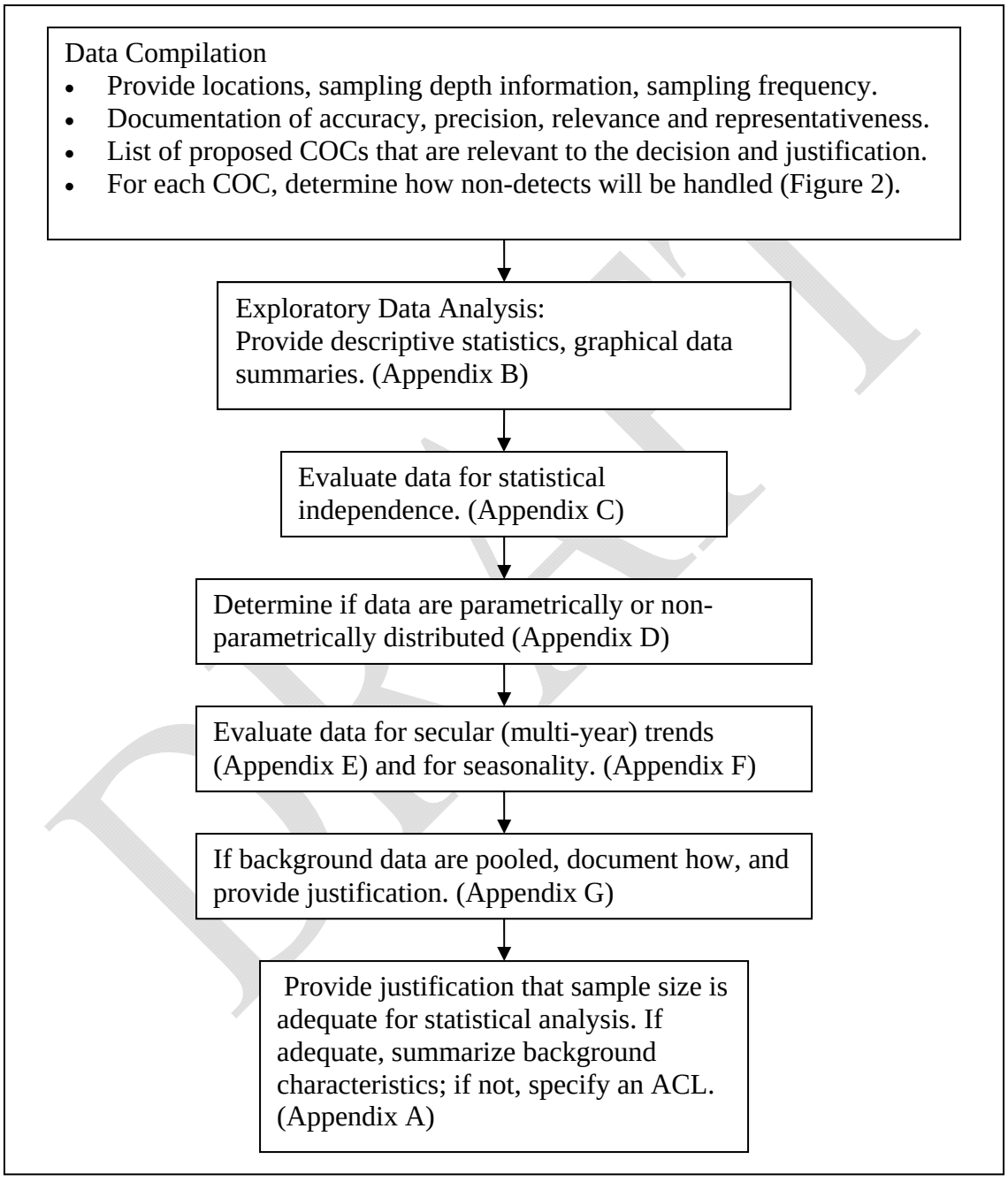


Figure 3. Procedure for evaluating background ground water quality and data adequacy

Appendices B through H provide some suggested methods to conduct the analyses shown in Figure 3.

IV. Statistical Determination of Water Quality Degradation

The term "degradation" is defined by the Ground Water Quality Rule "as the lowering of ground water quality as measured in a statistically significant and reproducible manner." To be statistically significant and reproducible suggests that multiple measurements over time are required before degradation can be established. The number of measurements and the length of time will likely be site-specific and dependent on the complexity of the situation.

Once background ground water quality is established, the next step is to determine the concentration at which an exceedance would be statistically significant and constitute degradation. Whatever statistical method is used, it shall provide a statistical decision threshold and a confidence level. Future downgradient measurements will be compared to this threshold to determine if degradation has occurred.

Some issues to be addressed when selecting an appropriate statistical decision threshold include:

- Are the data adequate to justify a statistically based decision threshold? (or would an ACL be more appropriate? see Appendix A)?
- Is the activity being evaluated new to the site? (this determines how future downgradient water quality will be evaluated to identify site impacts).
- Should interwell or intrawell comparison methods be used? (depends on whether upgradient and downgradient comparisons of wells are possible and defensible).
- Is a tolerance interval or a prediction interval appropriate and justified for the problem at hand?

IV.1 Alternative Concentration Limit

Alternative Concentration Limits (ACLs) for constituent(s) of concern are to be estimated when there are insufficient data to meet the statistical assumptions for a more detailed statistical analysis. An ACL is to be used as an interim upper regulatory threshold in order to be fully protective of human health and the environment in situations where sufficient data are lacking to adequately define background water quality and/or an appropriate statistically defensible upper threshold based on background is not available.

ACLs may also be used under certain circumstances, agreed to by all involved parties including DEQ, when a rigorous statistical evaluation is not desired, practical, or necessary. ACLs are to be established on a case-by-case basis in consultation with IDEQ. The ACL estimation process is described in Appendix A.

IV.2 New vs. Existing Activity

An appropriate statistical decision threshold should be chosen in consideration of whether an activity is new or existing. For example, the considerations recommended for wastewater land application sites are shown in Table 1.

Table 1. Considerations for Wastewater Land Application Sites

Facility with no previous site impact (new or existing)	Facility with existing site impacts
• Downgradient wells can also be used to define background ground water quality	• Only wells unaffected by the facility's operation can be used to define the background ground water quality
• Decisions are made via intrawell comparison	• Decisions are made via interwell comparison (or intrawell, if warranted)
• Upper tolerance limit (UTL) or Shewhart-CUSUM control chart limits are used set decision threshold	• Upper prediction limit (UPL) is used as a decision threshold
• Multiple downgradient wells compared to background UTL or individual well compared to its control chart limits	• Multiple downgradient wells compared to upgradient UPL; verify exceedance with two resamples

IV.3 Interwell vs. Intrawell Analysis

The objective of degradation analysis is to identify an appropriate background data set against which concentrations in wells potentially affected by a facility can be compared, so as to monitor the facility's impact on local water quality. Generally, interwell comparisons are appropriate where water quality is spatially homogeneous, sample locations provide statistically independent data, and appropriate upgradient - downgradient comparisons can be identified and defended. Intrawell comparisons are usually applied in wells where water quality has not been impacted by site activities and therefore represents background water quality at that location. However, intrawell comparisons may be preferable in certain situations where strong spatial variability exists or is impossible to assess using a single upgradient well (see Appendix N).

IV.4 Decision Thresholds and Confidence Levels

Decision thresholds that are commonly used in ground water monitoring analysis are the prediction limit, tolerance limit, and confidence limit. An **upper prediction limit** specifies the maximum allowed concentration that 100% of the next k comparisons must fall below in order to avoid an exceedance designation at a designated level of confidence (i.e. 95%) (EPA, 1992a); an **upper tolerance limit** specifies the upper limit that a designated percentage of all future comparisons (e.g. 95%) must fall below to avoid a designation of contaminated with a designated degree of confidence (e.g. 95%) (EPA, 1992a); and **a confidence interval** brackets the range of a specified population parameter (e.g. the mean) at a designated level of confidence (i.e. 95%) (EPA, 1992a). Simultaneous intervals on multiple constituents are beyond the scope of this guidance; we assume that

future comparisons are made relative to historical background data on a constituent-by-constituent basis.

Prediction and tolerance limits are applied for compliance sampling in detection, assessment, and monitoring programs since only one initial sample per well is required during the compliance period. They are also used for establishing background-based ground water protection standards. Confidence intervals are most often used when comparing water quality measurements to a ground water standard which is based on a mean or median value (Virginia DEQ, 2003). Before calculating these limits, it should be confirmed that the background data are statistically stationary, independent, and normally (or lognormally) distributed (see Section V for more details on these terms).

Some rules of thumb to keep in mind when considering confidence intervals are:

- Wider statistical intervals are associated with higher confidence levels. However, too high a confidence level decreases the power of the test (the probability of detecting an exceedance) so the confidence level ($1-\alpha$) should not be set higher than necessary. Conversely, too low of a confidence level may result in an excessive number of exceedances.
- The conservative choice when testing for a trend or a difference is to use a narrower interval or a lower confidence level (90% or 85%). This would reduce the probability that a difference or exceedance may be missed. In most cases a 95% confidence level ($\alpha = 0.05$) provides the best compromise between power and confidence.
- For non-parametric methods in which the confidence level is dependent on sample size, the highest confidence level for the available sample size is typically selected. Larger sample size may be needed to achieve a desired confidence level.

IV.4.1. Intrawell Tolerance Limits

At sites where ground water has not been previously affected by site activities, future water quality measurements will be compared to background water quality in the same well, via the intrawell upper tolerance limit(UTL). The UTL sets the background water quality for each constituent of concern in each monitoring well; it is described in more detail in Appendix H. Application of the method requires that the data are normally or lognormally distributed and have been corrected for seasonal effects or are temporally stationary. For data that meet the above requirements but are not normally or lognormally distributed, a non-parametric UTL can be calculated. Appendix I contains information on the sample sizes needed for non-parametric UTLs.

IV.4.2. Interwell Prediction Limits

In cases where site conditions indicate that the ground water quality in downgradient wells differs from background conditions (because of existing site practices), data from multiple downgradient wells will be compared to upgradient wells (an interwell analysis)

via a parametric upper prediction limit (UPL) calculated from pooled upgradient background data.

Application of the method requires that the data be normally or lognormally distributed and corrected for seasonal effects. For lognormal distributions, the UPL can be determined using the method outlined in Appendix J. For data that meet the above requirements but are not normally or lognormally distributed, a non-parametric PL can be calculated (see Appendix K).

In cases where background ground water quality data are non-stationary, the trend must be evaluated. Water quality may show a trend in response to 1) natural circumstances or 2) anthropogenic activities such as land use changes. Regardless of the cause, the trend may disappear over time. The methods described in Appendix L can be used to establish interim decision thresholds until a new stationary background is achieved.

Examples of application of the guidance to new and existing wastewater land application sites are provided in Appendix M and Appendix N, respectively.

V. Summary of Process

V.1 Determination of Background Ground Water Quality

New sites have the advantage that all monitoring wells, regardless if they are up-gradient or down-gradient, can be used as background monitoring wells. For example, at WLAP facilities where land is converting from another land use (such as irrigated agriculture) to wastewater application, it is possible that some or all wells will not have attained a steady state condition (see below) or that down-gradient wells will have reached a different steady state condition than up-gradient wells.

For a new site or new unused acreage at an existing site, such a WLAP facility that has yet to have any wastewater applied, the first step is to conduct descriptive statistics on the constituent(s) of concern in all background and compliance monitoring wells (Appendix B). Following the initial descriptive statistical tests, each of the monitoring wells should be evaluated for data independence (Appendix C). The form of the data distribution (parametric or non-parametric) should be determined next (Appendix D).

Statistically significant seasonal trends for each of the constituent(s) of concern (Appendix E) are then evaluated and such trends removed to produce a seasonally stationary data set. As the regulated entity is required to evaluate at least three years of quarterly data (where each quarter's data represents the same month from year to year), some of the background water quality variation may be due to changing land use practices (e.g. nearby agricultural activities, stream and canal flow, etc.) or climatic changes (precipitation patterns, evapotranspiration, etc.). The preferred method for determining seasonal stationarity is the non-parametric Kruskal-Wallis test (Appendix E).

The resulting data set should then be checked for the presence of secular (long-term) temporal trends (Appendix F). If a trend exists, then setting degradation thresholds may not be statistically valid and can lead to erroneous conclusions. The recommended method for testing for temporal stationarity is the non-parametric Mann-Kendall test for trend.

If the Mann-Kendall test shows that there is a statistically significant secular trend (either positive or negative), then an alternative method needs to be followed to set the standard(s) that the regulated entity will need to follow; see Section V.2.c. below. If the Mann-Kendall test reveals no secular trends, the regulated entity can proceed to determine whether the data from multiple background wells can be pooled (Appendix G).

V.2 Determination of Degradation

At this point, background water quality has been rigorously evaluated and its statistical characteristics identified. The next step is to define appropriate thresholds against which future measurements can be compared to identify potential water quality degradation.

V.2.a Intrawell Comparisons

Parametric tolerance levels for intrawell comparisons can be set using the methodology provided in Appendix H. An intrawell analysis compares future constituent levels in a well to the limit established by that well's own background water quality. In order to use this method, one must have a data set that (1) is stationary (free of secular trends and has no statistically significant seasonal effects or has been corrected for seasonality), (2) is normally or lognormally distributed, and (3) represents a site whose ground water has not been impacted by previous site activities. Appendix I provides a methodology for determining non-parametric tolerance limits (where the same assumptions apply). Future water quality for each COC in each well is to be compared to the upper tolerance limit in each well. If the rate of exceedances in future measurements exceeds that used to establish the limit (e.g., 5% of all future measurements), then the site is deemed to be out of compliance. Appendix N provides the Shewhart-CUSUM control chart method which monitors gradual and rapid site impacts.

V.2.b. Interwell Comparisons

The methodology for setting parametric prediction levels for interwell analyses is provided in Appendix J. In this case, an upper prediction limit is defined on the basis of up-gradient water quality data. In order to use this method, one must have a data set that (1) is stationary (free of secular temporal trends and has no statistically significant seasonal effects or has been corrected for seasonality), and (2) has been found to meet the parametric distribution assumptions. Site conditions must be such that the down-gradient well water quality can be compared to up-gradient water quality (an interwell analysis). Appendix K assumes the same conditions as Appendix J except the distribution is non-parametric. A specified number of future water quality measurements in down-gradient wells are compared to the upper prediction limit established in up-gradient wells; any exceedance is to be verified by the verification resampling procedure discussed in Appendix J.

V.2.c. Interim Methodology for Trending Data

Appendix M outlines a suggested procedure for setting an interim upper prediction limit for situations that violate the stationarity assumption (the data show a secular trend).

VI. Statistical Concepts

VI.1 Glossary and Description of Statistical Terms

Throughout this document, certain mathematical symbols will be reserved for quantities related to sample size such as the number of observations, number of years of sampling, and frequency of sampling within the year. Other symbols will be used to denote the sample mean, standard deviation, and other sample-based statistics. For reference, some of the frequently used symbols are summarized in Table 2.

Unless stated otherwise, the symbols $x_1, x_2, \ldots, x_N$ are used in this guidance to denote a chemical concentration measurement in each of N ground water samples taken at regular intervals during a specified period of time. The subscript indicates the order in which the sample was drawn (e.g., x_1 is the first or oldest measurement while x_N is the Nth or latest measurement). Collectively, the set of x's is referred to as a data set, and in general x_i will be used to denote the i^{th} measurement in the data set.

Table 2. Summary of Statistical Notation

Symbol	**Definition**
x_i	Constituent concentration measurement for the i^{th} ground water sample
m	The number of years for which data were collected (usually the analysis will be performed with at least 3 full years worth of data)
n	The number of sample measurements per year (for quarterly data, n=4). This is also referred to as the number of "seasons" per year.
N	The total number of sample measurements (for m complete years' worth of data, N = nm).
$\bar{x}$	The mean (or average) of the chemical measurements in a sample of size N.
s	The standard deviation of the chemical measurements in a sample of size N.
s^2	The variance of the chemical measurements in a sample of size N.

VI.2 Terminology/Definitions:

A **population** is the set of all possible measurements of interest in the real world. For example, an aquifer's nitrate concentration represents a population of all possible aliquots of water that could be collected from that aquifer and analyzed for nitrate.

A **sample** is a set of measurements collected from the population in a manner that attempts to be representative and unbiased. In the preceding example, a sample might

consist of 20 measurements (the **sample size**) collected at randomly chosen wells from across the aquifer.

A **parameter** is a numerical measure of the characteristic of the population being sampled. Typical parameters are the population mean (μ), variance (σ^2) or standard deviation (σ), and proportion (p). Parameter values are usually unknown.

An **estimate** is a numerical measure of a parameter derived from a sample such as its mean ($\bar{x}$), variance (s^2) or standard deviation (s), and proportion ($\hat{p}$).

Inference is the process applied to estimate a population's parameter(s) from the sample. The parameters are usually the targets of our interest but, because it is impossible or prohibitive to collect every measurement from the population, they are usually unknown. There are two approaches to making inferences: estimation and hypothesis testing. The first answers the question, "what is the value of the parameter?" and the second answers the question, "Does the parameter meet this specific value?" In groundwater analysis, the corresponding questions might be "what is the nitrate concentration in the groundwater?" or "does the nitrate concentration meet State standards?".

Data independence is the most basic requirements of statistical inference. All measured values in a sample are assumed to be random. In a time-series sense, a measurement must not depend on—or affect—any prior or future measurement. In a spatial sense, a measurement made in one well must be independent of those made in other wells. Data that violate this requirement (e.g., as in replicate measurements collected over a short time span in the same well) carry redundant information that biases the calculation and/or inference of any statistical quantity based on it. See Appendix C for further information.

The sample **distribution** is the frequency or probability of occurrence of the measured values. In groundwater analysis, two commonly encountered distributions are the **normal distribution** (bell-shaped curve) and the log-normal distribution (having the normal shape when values are logarithmically transformed). **Transformation** of the data to a normal distribution via a mathematical function is often necessary for environmental data, prior to statistical analysis. Estimates derived from a random sample, (e.g., $\bar{x}$) are also random; different samples generate different estimates. Estimates derived from different samples of the same population are known as the **Sampling Distribution**.

Many common statistical methods are based on a knowledge of, or the assumed characteristics of, the sampling distributions. One of the most famous is the **Central Limit Theorem** which says that the sampling distribution of the mean of many independent random samples is normal regardless of the underlying distribution of the population that was sampled.

VI.3 Methods for Describing a Distribution

Data need to be summarized in order to make meaningful interpretations and to pose testable hypotheses. Data summarization can be graphical or numerical. Graphical methods emphasize the shape of the distribution and numerical methods emphasize its central tendency and dispersion.

Graphic methods

Histogram summarizes the frequency distribution of a data set by displaying the number of observations that fall in defined intervals. Suggested numbers of intervals is 5-15, but the number of intervals can greatly affect the visual appearance of the distribution. The Y-axis can either be the number of occurrences in each interval or the percentage of total occurrences.

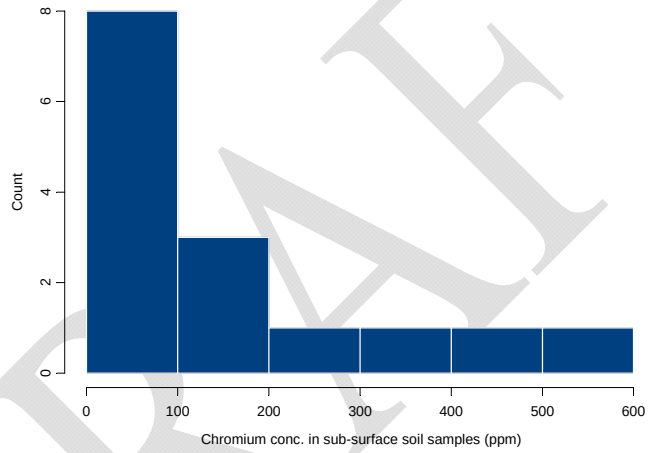


Figure 3. Example Histogram

Box plot is schematic representation of a distribution. The following values are summarized in a box plot: the box - the center 50% of the data (bounded by 25% and 75% of the observed values), the line in the box is the median, the whiskers are the upper and lower adjunct values. Generally, the upper adjunct value is the value 1.5 times the interquartile range (IQR) away from the 75% observed value and the lower adjunct value is the value 1.5 times IQR away from the 25% observed value. **Interquartile range** (IQR) is the difference between Q3 and Q1, showing the dispersion of the center-most 50% of the data. The IQR is robust to extreme values but cannot describe the overall nature of the dispersion.

Observations beyond the upper and lower adjunct values are extreme values; an extreme value in a data set is NOT necessarily a bad observation that needs to be removed. Box plots summarize a sample's central tendency, spread, skewness and extreme values. Side-by-side box plots are a useful tool for comparing the distributions of different samples or groups of measurements.

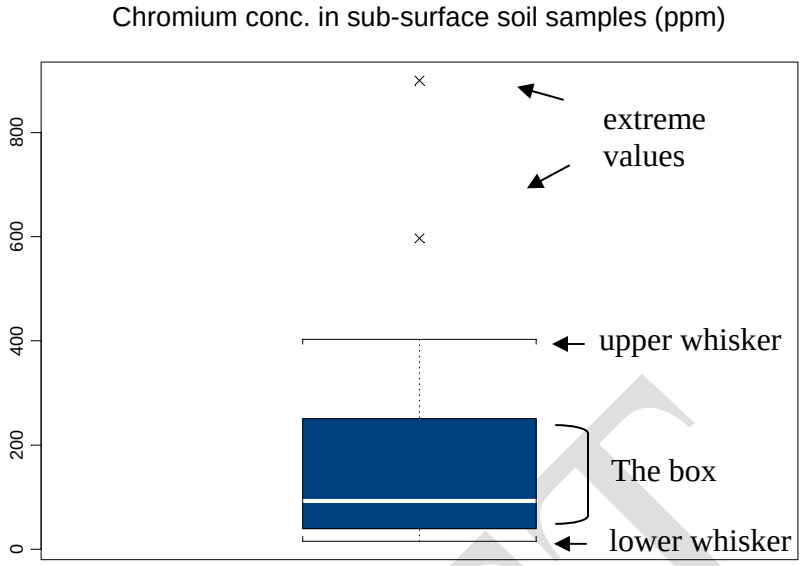


Figure 4. Example Box Plot

Time-series plot shows the values of observations on the y-axis vs. corresponding times they were collected on the x-axis. Equal-spaced time points are desirable. A time-series plot is useful for examining general trends over time, evaluating seasonal or cyclical patterns and disrupting events (such as the effect of a drought year on water quality). If the data points are not collected in equal time intervals, it is important to reflect the interval width between time points in the plot. Otherwise, the apparent visual trend could be misleading.

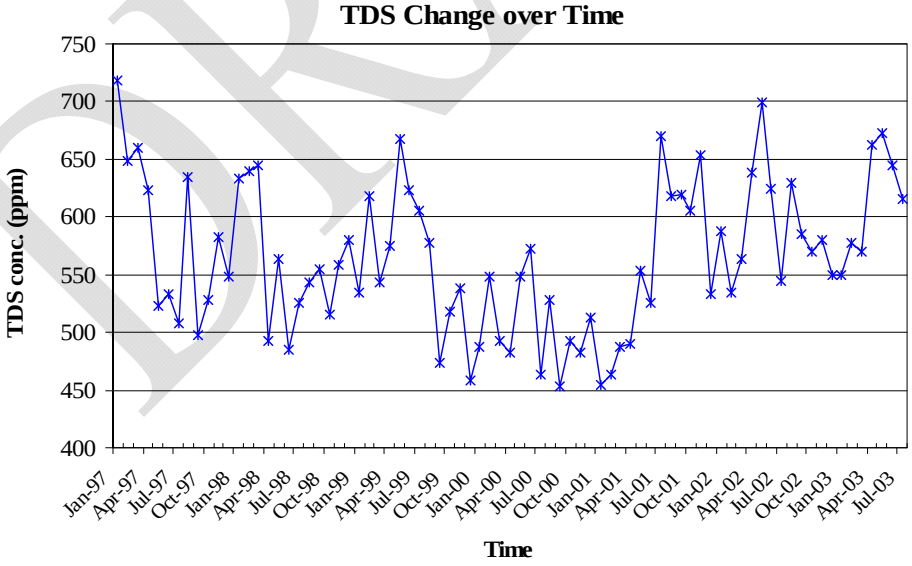


Figure 5. Example Time-series Plot

Scatter plot is used to examine the relationship between two variables, x and y. Each point on the scatter plot represents a pair of measurements of x and y from the same source, e.g., TDS and NO3-N concentration in the same well sample. Usually we are

interested in determining if there is a linear or non-linear correlation between the two variables.

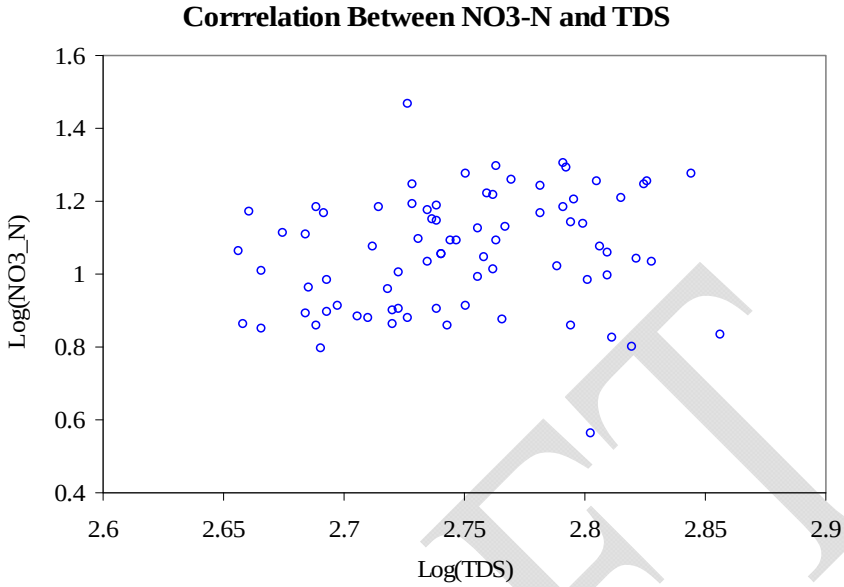


Figure 6. Example Scatter Plot

Numerical Methods

Central tendency is a distribution's "center of mass". Common measures of central tendency include the mode, median and mean. The **mode** is the most frequently occurring value in a data set. Distributions can have more than one mode (bimodal, trimodal, etc.).

Median is the middle value of a data set. It is the 50th percentile of a distribution, in which half of the observations are less, and half are greater, than the value. In a data set whose N observations are arranged from smallest to largest, the location of the median is (N+1)/2 from the bottom of the list (or the average of the two middle observations).

Mean or arithmetic mean or average is the sum of N observations divided by N. The goal of statistical inference is to estimate the population mean (μ) from the sample mean. The sample mean ($\bar{x}$) is calculated as:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{N}$$

where N is total number of observations in the sample. The mean is sensitive to extreme values in a given data set and therefore may not always represent a distribution's central tendency. The median is robust to extreme values and thus is a better measure of central tendency in skewed distributions. For a symmetric distribution, the mode, mean, and median are the same; for skewed distributions, they are different. The following graphs show their relationships in various distributions.

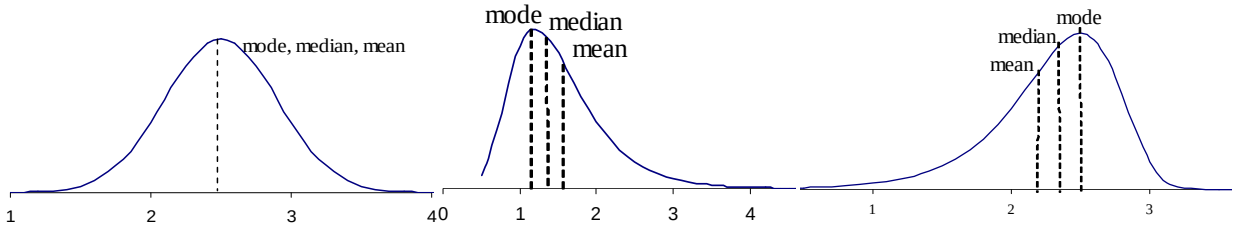


Figure 7. Mode, median, mean for various distributions

(from left to right: symmetric, positively skewed and negatively-skewed)

Dispersion is the **spread or variability** around the central tendency. Common measures include the range, interquartile range (IQR), variance and standard deviation. **Range** is the difference between the largest and smallest values in a data set. Although it is simple to calculated, it is least useful in describing dispersion since it reflects the extreme values. A better measure involves the use of **quantiles**. The p^{th} quantile of a data set is the value that *p* percent of the observations are less than or equal to. The most common quantiles are the 25^{th} (Q1 or first quartile), 50^{th} (Q2 or median) and 75^{th} (Q3 or third quartile).

The **variance** of a sample data set is the average of the squared deviations of the observations from the mean. It is calculated as:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{N-1}$$

The **standard deviation** (SD) is the square root of the variance, $s = \sqrt{s^2}$. For a normal distribution, the following empirical rule applies: 68% of the measurements are within one SD of the mean, 95% are within two SDs and 99% are within 3 SDs of the mean.

$\bar{x}$ and s or s^2 are the most commonly used descriptive statistics for a distribution's central tendency and dispersion. However, they are most appropriate for symmetric distributions because the are sensitive to extreme values. To adequately characterize a skewed distribution, the range, Q1, median, and Q3 should be reported.

Skewness is the third moment of a distribution and measures its asymmetry. Skewness is zero for a symmetric distribution and either **positively skewed** (skewed-to the-right) or **negatively skewed** (skewed-to-the-left) for asymmetric distributions. In a positively skewed distribution, the measurements tend to cluster around smaller values and tail toward larger values.

Kurtosis is the fourth moment of a distribution and measures the sharpness of its peak. Kurtosis for a normal distribution is equal to 3.0. A kurtosis greater than 3.0 (or zero in some statistical packages, which subtract 3 from calculated moment) indicates a distribution that is more sharply peaked than a normal distribution. The example in Figure 8 shows different example distributions, one with a kurtosis of >1 (red,), one with zero kurtosis (blue, a normal distribution) and one with a kurtosis of <0 (green).

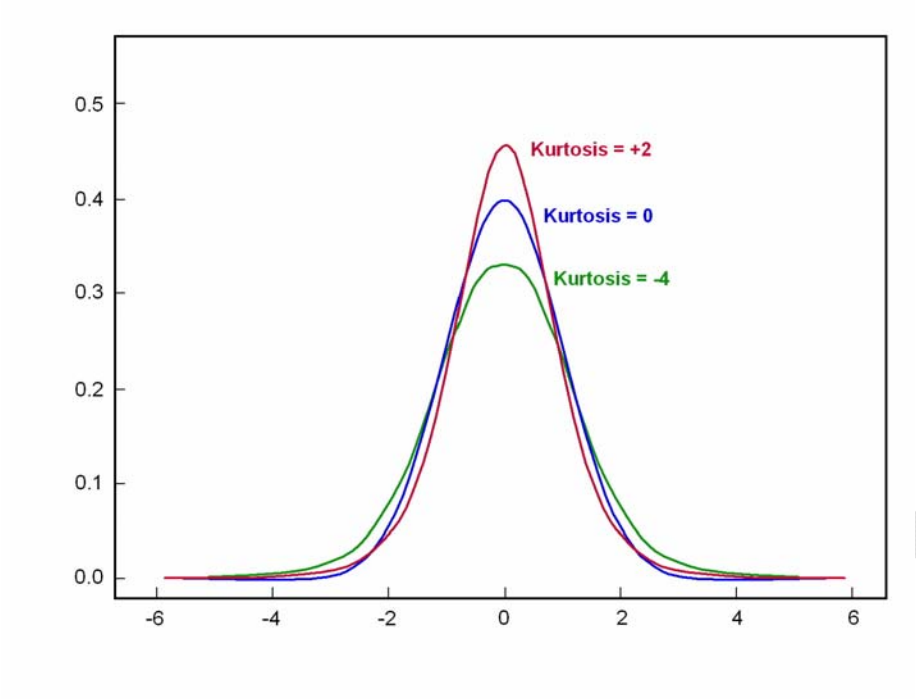


Figure 8. Example of some distributions with various degrees of kurtosis (peakedness).

VI.4 Inference: Estimating Decision Thresholds

Some decision thresholds that are commonly applied in ground water analysis are the prediction limit, tolerance limit, and confidence limit. An **upper prediction limit** is the value, based on existing measurements and a specified level of confidence, below which the next *k* future measurements are expected to fall. An **upper tolerance limit** is the value defined at a particular confidence level for a specified percentage of all future measurements**.** In contrast, a **confidence limit** defines a permissible range for a specified population parameter (e.g., the mean) at a specified level of confidence. All three limits are calculated from historical background data on a constituent-by-constituent base and are used to compare future measurements to determine whether the sampled population has changed (as by contamination).

Prediction and tolerance intervals are applied for compliance sampling events in detection, assessment, and monitoring programs and can be used for establishing background-based ground water protection standards (EPA, 1992a). Confidence intervals are often applied for comparing measurements to a ground water standard which is based on a mean or median value (Virginia DEQ, 2003). Before such intervals are calculated, the background sample distribution should be checked for normality or lognormality, stationarity and data independence.

Upper prediction limit (UPL):

$$\text{UPL} = \bar{x} + Ks$$

Where *K* is the one-sided normal tolerance factor that can be found tabulated in various ground water monitoring guidance documents (Gibbons, 1994; EPA, 1992a). As Gibbons (1994) has pointed out, *K* must be calculated for a specified statistical model which includes a verification sampling protocol, background sample size, and *k*, the number of relevant[1] future measurements that will be compared within a specified time period). See Appendix J for an example table of *K* factors. If any constituent of concern exceeds the UPL during the comparison time period, then further sampling according to the verification sampling protocol is conducted to verify the exceedance.

Upper tolerance limit (UTL):

$$\text{UTL} = \bar{x} + Ks$$

Where *K* is the one-sided normal tolerance factor defined for a specified fraction (the "coverage", e.g., 95%) of all future comparisons (Gibbons, 1994; EPA, 1992a). A specified number of exceedances are allowed as long as their total number is no greater than the specified percentage of comparisons made since the UTL was set (e.g., 5% for 95% coverage.

Upper confidence limit (UCL):

$$\text{UCL} = \bar{x} + t_{n-1,\alpha}\frac{s}{\sqrt{n}}$$

where $\bar{x}$, *n* and *s* are the average, number and standard deviation of the background data, respectively, and t is the t-statistic with n-1 degree of freedom at a 1-α upper-tail confidence.

If the data are log-normally distributed, all of the above limits should be calculated on a log-transformed scale and compared with data that have also been log-transformed.

VI.5 Inference: Hypothesis testing

The second common type of statistical inference is aimed at answering the question, "Does the population parameter meet *a specific condition or value*?" For example, does the mean NO3-N concentration of the upgradient wells around a land application facility exceed the 10 mg/L Idaho Ground Water Quality standard? The question is assessed by examining the sample characteristics relative to a statistical hypothesis concerning the population's characteristics.

A statistical hypothesis is a statement about a parameter in a population and a **hypothesis test** is a formal procedure for comparing the sample data with a hypothesis whose truth we want to assess (Moore and McCabe, 1998). The results of a test are expressed in terms of a probability that expresses how well the hypothesis agrees with the data.

[1] Any measurement that is deemed useful for detecting the monitored facility's impact. The number of future comparisons is defined on the basis of number of constituents of concern examined per well, number of wells, sampling frequency and duration of the comparison time period. See Appendix J for details.

A hypothesis test involves four steps:

1) State the hypotheses and confidence level: a null hypothesis (H_0) is the statement being tested; an alternative hypothesis (H_A) is the statement we will accept should H_0 be rejected. The significance of the test, α, is the compliment of the confidence level ($1-\alpha$) and indicates the strength of the evidence against the null hypothesis. The smaller the α, the less the chance of falsely rejecting H_0.

2) Choose and compute the test statistic. A test statistic provides a quantifiable measure for deciding between H_0 and H_A. Some examples are the t statistic and the F statistic.

3) Find a p-value based on the test statistic. The p-value is the lowest significance level at which H_0 can be rejected (or the probability of obtaining a test statistic as extreme as or more extreme than that calculated from the sample, if H_0 were true).

4) State the conclusions based on a decision rule: 1) if the p-value is less than α, then reject H_0 and accept H_A; 2) if the p-value is greater than or equal to α, then we cannot reject H_0 based on information provided in the data set. Both decisions are made at an α significance level ($1-\alpha$ confidence level).

Two types of errors are associated with any hypothesis test, A Type I (false-positive) error occurs when falsely rejecting H_0; a Type II (false-negative) error occurs when falsely accepting H_0. For example, if the null hypothesis, H_0, asserts that ground water is not contaminated, a Type I error is to claim that contamination exists when it actually does not. A Type II error would claim that ground water is not contaminated when it actually is. The risks of committing the two types of errors are α and β, respectively and they are complimentary: specifying a low value of α means accepting a high β; $\alpha = 0.05$ is usually considered an acceptable trade-off between the two risks. Just as ($1-\alpha$) is the confidence level of a hypothesis test; ($1-\beta$) is the **power** of the test (the likelihood of identifying contamination if it is present). Type II errors are more likely for small sample size, so β should be considered at the time of sampling design. Sample size should be large enough to detect differences in population parameters with a power of at least 80% and will vary depending on the confidence level.

VI.6 Sample size

Sample size affects both estimation and hypothesis testing. For estimation, it determines the estimate's precision, and for hypothesis testing, it affects the power of the test. Sample size depends on the type of statistical test chosen and also on measurement precision. A large sample size is almost always desirable. Having many observations can make trivial differences detectable. The goal of determining sample size in a statistical study is to find the number of samples which provides adequate yet practically feasible evidence to draw meaningful conclusions relative to the goals of the study. It is always

good practice to state the problem first and then set up decision rules to address the problem.

For ground water analysis, at least twelve background samples must be available for determining decision thresholds and for making interwell comparisons. The samples must be statistically independent and representative of seasonal and spatial variability at the site. For this reason, the eight samples preferably should be collected quarterly over a two year period in a well. For interwell comparisons with two upgradient wells reflecting statistically indistinguishable chemistry, one year of quarterly data for each well is required (if the two wells' chemistries are different, then two years of quarterly data at each well should be available). Statistical analysis can be conducted with smaller data sets, but smaller sample size usually leads to such wide prediction intervals that no meaningful conclusions can be drawn. The statistical requirements of the various analysis methods should be understood so adequate numbers of samples are collected prior to analyzing the data.

VI.7 Non-parametric methods

Data sets having a normal or log-normal distribution should be analyzed with parametric methods. Parametric methods are more powerful because the actual values of the measurements are used in the analysis. Parametric methods assume some knowledge of the shape of the distribution (i.e. normal or lognormal) and use the measured data values to estimate population parameters. For example, the t-test is a parametric method for bell-shaped distributions (either in original scale or transformed scale) that are centered at μ with a dispersion of σ.

Sample distributions that do not have a normal or log-normal form, can be analyzed with non-parametric methods that don't require assumptions about the form of the population distribution. The only requirement is that the population distribution be continuously valued; additionally, if two populations are to be compared, then they should have the same shape.

It is common that sample size is inadequate to determine whether a particular distribution is parametric, or that the number of non-detects is too large to determine the form of the distribution. In such cases, non-parametric methods can be used both for establishing background levels and for hypothesis testing. These methods do not require assumptions about the form of the distribution and can be used to estimate parameters or to test a hypothesis.

Some common non-parametric methods used in ground water analysis are based on ranks of the data but not the actual values. Data values are ordered from lowest to highest and ranked according to their position in the ordered list. Commonly used rank analysis methods include the **Wilcoxon signed rank** (non-parametric equivalent of the one-sample t-test), **Wilcoxon rank sum test or Mann-Whitney's test** (non-parametric equivalent of the two-sample t-test), **Kruskal-Wallis** (non-parametric form of the multiple-sample ANOVA test) and **non-parametric regression**. As for parametric

methods, statistical independence of the observations is required for all non-parametric methods.

The **Kruskal-Wallis test for seasonality** can be used to test for the presence of significant seasonal fluctuations in a time-series data set. **Mann-Kendall's test** for trend shows if a significant secular trend exists, and **Sen's test** estimates the slope of the trend, regardless of the presence of missing observations or variable sampling time intervals.

Non-parametric prediction and tolerance limits are based on the maximum values observed in N background measurements, where sample size depends on confidence level and future comparison strategy (Appendix I and K). Confidence level in turn is a function of the number of future comparisons (k) and the exceedance verification sampling plan. These methods require very large background sample sizes if k is large or if α is small, so that trade-offs are usually required.

Bootstrap and Jacknife resampling methods are recently developed non-parametric methods for making statistical inference. Basically, the original sample data set is randomly resampled thousands of times and statistics of interest recomputed each time. The calculated statistics from all resampled data sets are used to estimate the relevant sampling distributions. In the Bootstrap method, resampling is conducted with replacement (of size N, the original sample size). In the Jacknife method, resampling systematically leaves out one value from the original data set each time (sample size = N-1). Unfortunately, small sample size is a major limitation because the resampling method assumes that the original data set is representative of the underlying population. These methods are computer-intensive but demonstrate growing potential for environmental statistical analysis. Their technical aspects are beyond the scope of this document. IDEQ leaves it to the regulated entity to choose the methods that best fulfill the objectives of the statistical analysis but retains the right to ask for alternative methods if they prove to be more appropriate.

Appendices

Appendix A. Alternative Concentration Limits

Alternative Concentration Limits (ACLs) for constituent(s) of concern are estimated when there are insufficient data to meet the statistical assumptions for a more detailed statistical analysis.

The following three measures of upper concentration limits are calculated from available data.

- $ACL_1 =$ the largest of the 12 most recent data values collected
- $ACL_2 = mean + 1.65s$
- $ACL_3 = \text{median} + 1.65 * \text{IQR}$ (where IQR = the interquartile range)

IDEQ specifies that the lowest of these limits is then to be used as an interim upper regulatory threshold in order to be fully protective of human health and the environment in situations where sufficient data are lacking to adequately define background water quality and/or an appropriate statistically defensible upper threshold based on background is not available.

ACLs are to be established on a case-by-case basis in consultation with IDEQ.

Appendix B. Exploratory Data Analysis/Descriptive Statistics

B.1 Descriptive Statistics

Once adequate data have been collected, the data should be analyzed using descriptive statistics to describe the overall population. At a minimum, the regulated entity should calculate the mean, standard deviation, skewness, median, minimum, and maximum for each constituent at each monitoring well and summarize the sample size for each constituent. In addition, the regulated entity should produce a visual representation of the distribution of each constituent (e.g., box plots or histograms) and concentration versus time plots for each well and each constituent.

As has been previously stated, the reason for collecting data is to try to understand the distribution of the true population. The sample mean provides a measure of the central tendency of the population, whereas the sample standard deviation provides a measure of its spread, or dispersion. The measurements represent just one of many possible subsets of data that could have been collected from the entire population. Different samples will obviously lead to different values of the sample mean and sample standard deviation. These differences are the reason why statistical intervals are used to infer population parameters and set decision thresholds.

In any set of data, it is possible that there will be outliers. Outliers can have one of three causes: (1) a measurement or recording error, (2) an observation from a different population, or (3) a rare event from the tail of the population of interest. If an outlier's cause cannot be detected or corrected, it should not be discarded from the data set.

Summary statistics can be calculated using any appropriate software (e.g., SPSS 2000's SysStat package; Microsoft's Excel, etc.). See Table B.1 for an example.

Other descriptive statistics include the median, minimum, maximum, and quartiles. The median and quartiles are not affected by outliers unlike the sample mean, standard deviation, and skewness.

A graphical summary of the data, including the relevant COCs, should provide box plots, showing at least the median, minimum, maximum, and quartiles for each constituent of concern and time-series plots. The latter provide a visual indication of whether there is a seasonal component to the data, whether there is a secular (long-term) trend, and/or whether the trend has changed or may be changing (approaching a new steady-state condition).

B.2 Example

Where appropriate, data from the following scenario will be used to illustrate selected applications in the various appendices of this guidance document. A wastewater land application facility wants to determine a background water quality level for TDS, above which there is a certain degree of statistical confidence that values would indicate

impacts to ground water. The facility has two background wells (#B1 and #B2). Well #B1 is located near an irrigation canal and the canal may seasonally influence the water quality. Well #B2 is located away from the irrigation canal. Three years of quarterly data have been collected at each monitoring well. Table B.1 contains the TDS data in parts per million (ppm) and an example of the summary statistics for TDS.

Table B.1. Data (parts per million) and resulting descriptive statistics for example scenario.

Time Index	**TDS Well #B1**	**TDS Well #B2**
Year 1 1st quarter	305	252
Year 1 2nd quarter	228	251
Year 1 3rd quarter	258	245
Year 1 4th quarter	259	252
Year 2 1st quarter	285	260
Year 2 2nd quarter	210	248
Year 2 3rd quarter	274	275
Year 2 4th quarter	240	272
Year 3 1st quarter	290	256
Year 3 2nd quarter	216	246
Year 3 3rd quarter	248	218
Year 3 4th quarter	235	225
Descriptive Statistics[1]		
Mean	254	250
Variance	904	268
Standard Deviation	30.0	16.4
Skewness	0.20	-0.52
Minimum	210	218
Maximum	305	275
Median	253	252
1st quartile	231	246
3rd quartile	290	258

[1] Excel software used for calculations

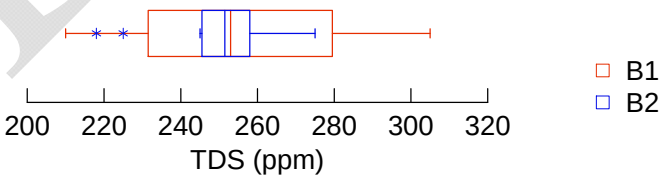


Figure B.1 Box plots for example data (from systat 10)

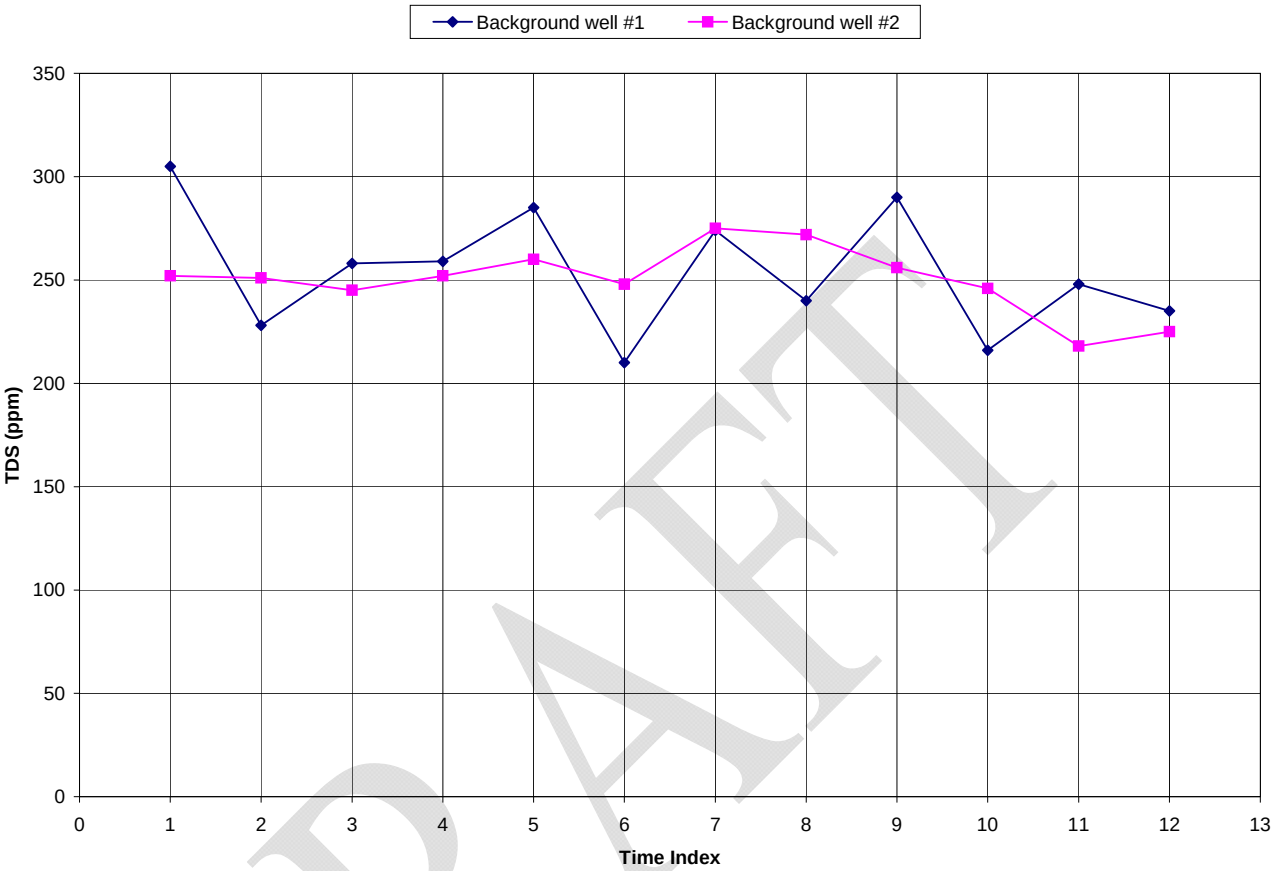


Figure B.2 Time versus concentration graph for example

Appendix C. Data Independence

C.1. Introduction and Background

All of statistical theory and practice is based on three fundamental premises.

1) First, every collection of measurement represents a purely random sample of the underlying population that is free of any bias imposed by considerations such as the sampling process, the name of the individual who conducted the sampling, or the analytical method used to make the measurements.

2) Second, the statistics of multiple samples collected from the same population are expected to fluctuate because of randomness; however, these fluctuations are assumed to bracket the underlying population statistics, allowing us to infer them from any sample (a statistical property known as ergodicity).

3) Third, and perhaps most important from a practical statistical standpoint, data are assumed to be independent, that is each measurement is randomly representative of the target population and independent of any other measurement. Dependent samples exhibit less variability resulting in the underestimation of the sample variance, which in turn affects the calculation of prediction limits and tolerance limits.

In reality, every measurement of the physical world possesses some degree of dependence on (similarity to, correlation with) previous or nearby measurements; this dependence is known as autocorrelation. For example, replicate measurements of stream chemistry are much more similar to (dependent on) each other than measurements collected a year apart. In another example, when drawing water from five different wells (n = 5), four aliquots of water from a single well are analyzed for quality control purposes. When calculating the average nitrate concentration in the five wells, the replicates cannot be treated as separate, random outcomes in a sample of n = 9 measurements, because they constitute redundant information about the population. If nitrate in that one well happens to be twice the average concentration of the other four wells, and we do average all 9 measurements together, then the apparent mean nitrate concentration would be biased high by 50% and the apparent variance would be far lower than the actual sample variance.

To avoid such bias and the ensuing decrease in the power of hypothesis tests and decision thresholds that it entails, every statistical procedure in this document assumes that the data analyzed are independent. Unfortunately, this is an area in which little guidance is available. For temporal data, two different approaches have been taken: one demonstrating physical independence based on the minimum time required for water to move past a well (EPA, 1989), and another, demonstrating statistical independence on the basis of historical time-series data (Barcelona et al., 1989; Oswina et al., 1992; Johnson et al., 1996; Ridley and MacQueen, 2005).

C.2. Example: Evaluating Data for Temporal Independence

Significantly more attention has been given to the issue of temporal autocorrelation than to spatial autocorrelation. The basic requirement is that sufficient time is allowed between sampling events to assure independence between samples. This can be evaluated with standard time-series analysis methods, but where sample size is too small ($N < 12$) for a time-series analysis, a method proposed by Ridley and MacQueen (2005), based on decision-tree logic and changes in concentration, could be adopted as an interim guide to determining minimum sampling frequency until sufficient data are collected to perform a statistical time-series analysis. Where data have been collected on a monthly or more frequent schedule, the goal of time-series analysis is to determine data properties by adjusting the temporal dependence and determine the sampling frequency that ensures data independence. In most cases, this will not be possible given the availability and frequency of measurements in most monitoring campaigns. In the absence of appropriate historical data, a general rule of thumb is that groundwater quality data should not be collected more frequently than quarterly (Gibbons, 1994, p. 163, 185) and if replicate analyses or more frequent time-series data have been collected, that the average of replicates (Washington State, 2005) or the average of multiple measurements collected within a quarterly span be used as the quarterly value.

Where appropriate data are available and are regularly spaced in time, autoregression analysis can be performed in most common statistical software (e.g., Box-Jenkins autocorrelation plots). For irregularly spaced data, a one-dimensional semivariogram (Oswina et al., 1992) can be computed using various geostatistical software packages. However, if the sampling frequency of the available data is no better than quarterly then in most cases there is little or no benefit to be gained from such an exercise. As an example, the TDS data in Figure C.1 represent varying sampling intervals over a 3-year period, with an indication that monthly measurements tend to be fairly similar (autocorrelated). A Box-Jenkins analysis is inappropriate in this case, unless it is conducted only on evenly-spaced quarterly data. The resulting autocorrelation plot is shown in Figure C.2. The autocorrelation statistic is a function of lag (separation in time); it decays rapidly from zero-lag where autocovariance equals the population variance to a value near 0.0 at a lag of 180 days. This suggests that TDS measurements are statistically independent if collected at a frequency no greater than once every 180 days. In this case, the conclusion is erroneous because monthly measurements were not considered.

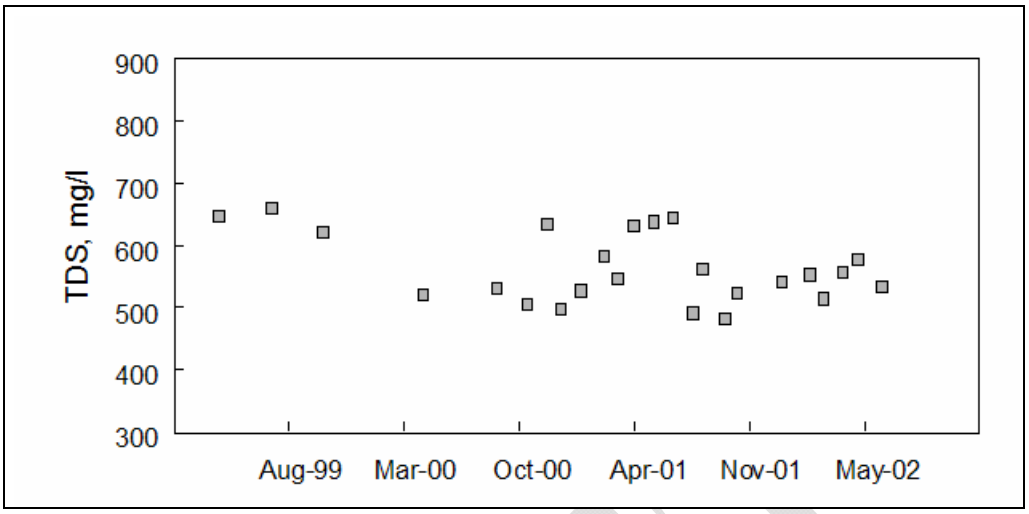


Figure C.1. Example groundwater TDS measurements for evaluating statistical independence of time-series data.

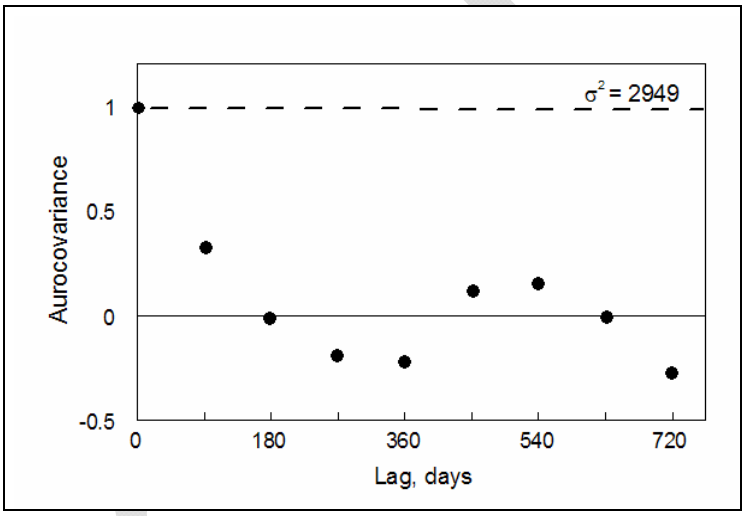


Figure C.2. A Box-Jenkins autocorrelation function plot

This Box-Jenkins autocorrelation function plot was created for the quarterly spaced TDS data in Figure C.1. The value of the autocovariance statistic decays to zero at a lag of about 180 days, suggesting that measurements spaced two quarters apart are statistically independent.

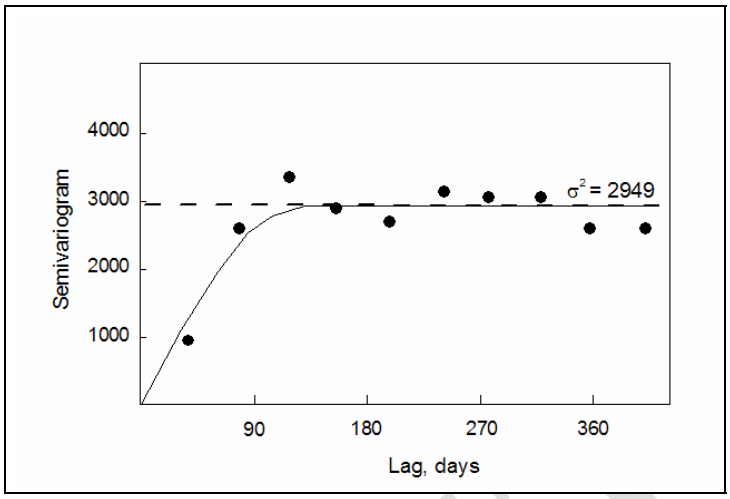


Figure C.3. Semivariogram

The semivariogram statistic (Isaaks and Srivastava, 1989) computed for all of the TDS data in Figure C.1, showing that the time-series data are statistically independent (pair-wise variance = population variance) if measurements are made at least 90 days apart (i.e., quarterly).

In order to determine the true minimum sampling interval allowed by the historical record, all of the data, including monthly measurements, must be considered. For irregularly sampled data, a one-dimensional semivariogram analysis can be performed with most geostatistical software (Oswina et al., 1992; Johnson et al., 1996); such a plot, utilizing all of the data in Figure C.1, is shown in Figure C.3. The value of the semivariogram statistic rises rapidly to a sill value that is at or near the population variance. The lag at which it achieves the sill (approximately 90 days) represents the minimum time interval (quarterly) for measurements to be considered statistically independent.

In this case, the semivariogram's temporal resolution is better than the Box-Jenkins plot because the data set contained measurements that were collected on a much shorter time interval than the Box-Jenkins plot could resolve with only quarterly data.

C.3. Evaluating Data for Spatial Independence

If existing guidance for evaluating temporal data independence is minimal, it is almost completely absent for spatial evaluations. Ideally, geostatistical data analysis could be applied to two-dimensional data sets in the same way that the one-dimensional semivariogram of Figure C.3 was applied to temporal data, the difference being that lags are defined in a spatial rather than a temporal sense.

As in the temporal situation, two approaches are possible for estimating the minimum spatial scale over which groundwater data can be considered independent: one based on physical considerations and another based on spatial statistical analysis (e.g., Bertolino et

al., 1983; Cameron and Hunter, 2002; Manly and Mackenzie, 2003). Such methods have limited applicability in small sampling networks (20-30 wells), however, and rarely are useful for small networks (< 5-10 wells). Figures C.4 and C.5 illustrate this point conceptually. Both the number and spacing of historic monitoring locations (as well as the type and availability of data at each well) determine how useful the geostatistical analysis can be.

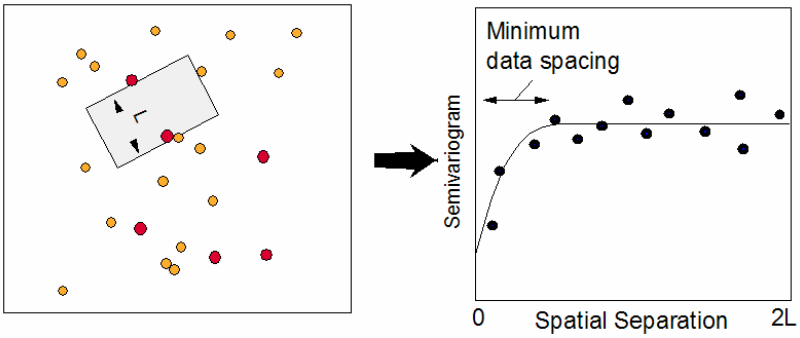


Figure C.4. Semivariogram with sufficient data

Hypothetical monitoring network in which there is a sufficient number of monitoring wells whose water quality data can be used to construct well-defined semivariogram statistics to identify the minimum well spacing for independent data.

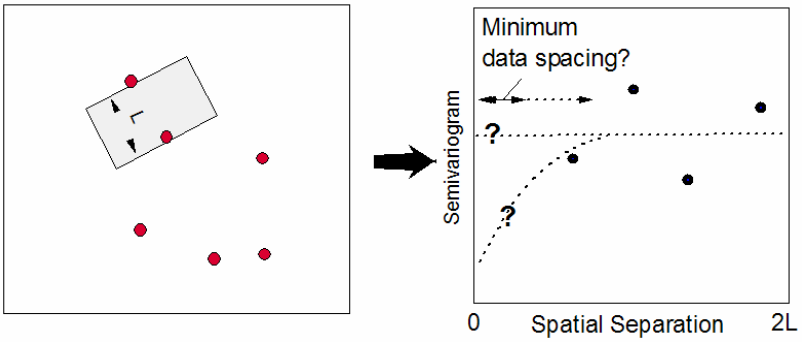


Figure C.5. Semivariogram with insufficient data

In almost all small monitoring networks there are insufficient wells to calculate reliable semivariogram statistics. Therefore, the minimum well spacing required for data independence cannot be determined.

As Gibbons (1994) and others have pointed out, the spatial hydrochemical variability of the site is as important as interwell spacing considerations. That is, if the aquifer is highly heterogeneous, then the assumption of spatial data independence may be violated for a different reason: the physical bias imposed on contaminant concentrations in active vs. inactive flow zones. That is, a well completed in a hydraulically "tight" zone will tend to reflect lower values of contaminant concentration than wells in more hydraulically active zones. It is for such reasons that intrawell evaluation methods may be the only rational alternative in highly heterogeneous aquifers (Gibbons, 1994).

In evaluating data independence in a spatial sense, IDEQ suggests foregoing a purely statistical evaluation of data independence in favor of a qualitative assessment, including but not limited to:

a) estimation of groundwater flow velocity between wells; in general, minimum permissible spacing tends to increase with groundwater flow velocity and with dispersivity;
b) examining the available data for concentration trends across the facility; if present, a spatial trend can strengthen autocorrelation in the direction of the trend and weaken the assumption of spatial independence in that direction;
c) an evaluation of overall site variability; if data from multiple upgradient wells cannot be pooled due to hydrochemical variability and if considerable hydraulic heterogeneity is known or suspected, then intrawell methods should be adopted if at all possible.

In the event that large spatial variability exists and/or the statistical independence of pooled upgradient data is suspect, intrawell analysis options should be explored. IDEQ may grant site-specific variances for intrawell analysis, utilizing modifications of the methods contained in this document or other methods such as those discussed in Appendix N.

Appendix D. Determination of Normality

D.1 Testing for Normality Using the Shapiro-Wilk Test

The primary reason to test whether data follow a normal distribution is to determine whether or not parametric test procedures can be employed. This is especially true when using tolerance intervals (EPA, 1988; EPA, 1992b). The null hypothesis (H_0) for any test of normality is that the data are normally distributed. Failure to reject H_0 does not prove that the data do follow a normal distribution, especially for small sample sizes, only that normality cannot be rejected with the evidence available. Use of a significance level (α) greater than 0.05 will increase the power to detect non-normality, especially for small sample sizes (Helsel and Hirsch, 1995).

The method described below is known as the Shapiro-Wilk Test for Normality. It is used for data sets with less than 50 data points. This method is recommended (EPA, 1988; EPA, 1992b; Fisher and Potter, 1989) as it is superior to the chi-square test (EPA, 1992a) and because it is based on probability plots (Helsel and Hirsch, 1995). The Shapiro-Wilk test is designed for data with less than 10-20% censoring, censored measurements up to this limit should be withheld from the calculation. If censoring is greater than 20%, then either Royston's method (Royston; 1993) or an appropriate adjustment to the sample standard deviation (Cohen, 1991. Aitchison, 1955) must be applied when using the formulae, below. The test is based on the premise that if the sets of data are normally distributed, then the ordered values should be highly correlated with corresponding quantiles taken from a normal distribution (Shapiro and Wilk, 1965). The Shapiro-Wilk statistic "W" is proportional to the ratio of the squared slope of the normal probability plot to the usual mean square estimate (Gibbons, 1994):

$$W = \frac{\left(\sum_{i=1}^{n} a_{i,n} x_{(i)}\right)^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}.$$

Alternatively, the formulas and example given below can be used as a guide (taken from EPA, 1992a).

D.2. Example

Consider the TDS data for well #B1 (Appendix B). The null hypothesis is H_0: distribution is normal. The coefficients $a_{i,n}$ for the W statistic are given in Table D.2 (see Gibbons, 1994 for a more complete table). Recalling that *s* is the standard deviation, W can be re-expressed as:

$$W = \left[\frac{b}{s\sqrt{n-1}}\right]^2$$

$$\text{where } b = \sum_{i=1}^{k} a_{n-i+1}\left(x_{(n-i+1)} - x_i\right) = \sum_{i=1}^{k} b_i$$

The general procedure is:

Step 1. Order the data from smallest to largest and list, as in Table D.1. Also list the data in reverse order alongside the first column.

Step 2. Compute the differences $x_{(n-i+1)}$-x_i in column 3 of Table D.1 by subtracting column 1 from column 2.

Step 3. Compute k as the greatest integer less than or equal to n/2. k=(n-1)/2 if n is odd and k=n/2 if n is even. Since n=12, k=6 in this example.

Step 4. Look up the coefficients a_{n-i+1} from Table D.2 and list in column 4. Multiply the differences in column 3 by the coefficients in column 4 and add the first k products to get the quantity b.

Step 5. Compute the standard deviation of the sample (9.77) and calculate W (0.861).

Step 6. Compare the computed value of W to the 5% critical value, equivalent to an α value of 0.05 (see Table D-3) for a sample size of 12 (0.859).

Table D.1 Example of Shapiro-Wilk Test for Normality on TDS Data from Well #B1

Column	Column 1	Column 2	Column 3	Column 4	Column 5
Ranked data value	x_i	$x_{(n-i+1)}$	$x_{(n-i+1)}$-x_i	a_{n-i+1}	b_I
1	242	268	26	0.5475	14.4175
2	244	268	24	0.3325	7.869167
3	246	266	20	0.2347	4.694
4	246	264	18	0.1586	2.8548
5	249	252	3	0.0922	0.245867
6	251	252	1	0.0303	0.0404
7	252	251	-1		b=30.12
8	252	249	-3		
9	264	246	-18	Std_dev =	9.77
10	266	246	-20		
11	268	244	-24	W =	0.8612
12	268	242	-26		

The closer the value of W is to 1.0, the greater is the support for the normality assumption. The assumption of normality is rejected if the computed value of W is less than W's critical value in Table D.3. In this example, the null hypothesis is accepted because W (0.861) is greater than the critical value (0.859). Therefore, the data are normally distributed.

Table D.2 Partial List of Coefficients a_i for the Shapiro-Wilk Test of Normality

# of data	8	9	10	11	12	13	14	15	16
k									
1	0.6052	0.5888	0.5739	0.5601	0.5475	0.5359	0.5251	0.5150	0.5056

2	0.3031	0.3244	0.3291	0.3315	0.3325	0.3325	0.3318	0.3306	0.3290
3	0.1743	0.1976	0.2141	0.2260	0.2347	0.2412	0.2460	0.2495	0.2521
4	0.0561	0.0947	0.1224	0.1429	0.1586	0.1707	0.1802	0.1878	0.1939
5		0.0000	0.0399	0.0695	0.0922	0.1099	0.1240	0.1353	0.1447
6				0.0000	0.0303	0.0539	0.0727	0.0880	0.1005
7						0.0000	0.0240	0.0433	0.0593
8								0.0000	0.0196

(Complete tables in Shapiro and Wilk, 1965; EPA, 1992a; Gibbons, 1994)

Table D.3 Lower 1% and 5% Critical Values for Shapiro-Wilk Test Statistic W

Sample Size	1% W Value	5% W Value	Sample Size	1% W Value	5% W Value
8	0.749	0.818	13	0.814	0.866
9	0.764	0.829	14	0.825	0.874
10	0.781	0.842	15	0.835	0.881
11	0.792	0.850	16	0.844	0.887
12	0.805	0.859			

(Complete tables in the following references: Shapiro and Wilk, 1965; EPA, 1992a; Gibbons, 1994)

The process for testing for lognormally distributed data is the same, except that the data are log-transformed prior to performing the Shapiro-Wilk hypothesis test.

Appendix E. Seasonal Trends

E.1 Testing for Seasonality Using the Kruskal-Wallis Test

One of the important requirements for conducting statistical tests is temporal stationarity, specifically; do the data exhibit seasonal variations in concentration? Note that k, the number of seasons, is defined to be appropriate for the data being analyzed; e.g., hourly stream temperature measurements might be grouped into two 12-hour "seasons" per day whereas monthly groundwater measurements are usually grouped into four 3-month "seasons." For measurements collected quarterly over a multi-year period (each quarter tested in the same month), some of the variation in background water quality may be due to changing land uses (nearby agricultural practices, river and canal flows, etc.) which can obscure seasonal variations in water quality due to precipitation, evapotranspiration, etc.

The Kruskal-Wallis test for seasonality is described below (taken from Gilbert, 1987; Helsel and Hirsch, 1995). This test is considered a non-parametric test, which means that the underlying population distribution is not assumed. The Kruskal-Wallis test may be computed by an exact method used for small samples sizes (see Lehmann, 1998 or Conover, 1999), or by a large-sample or chi-square approximation (Helsel and Hirsch, 1995). The null and alternative hypotheses are:

> H_0: All of the seasonally-grouped subsets of data have identical distributions
> H_A: At least one group differs in its distribution.

In other words, do the measurements taken in one quarter of the year differ significantly from the measurements taken in any other quarter of the year?

To conduct the test, the data are ranked from smallest to largest, from 1 to N. If H_0 is true, the average rank for each of the k seasonal groups should be similar and also be close to the overall average of the N data. When H_A is true, the average rank for some of the groups will differ from others, reflecting the difference in magnitude of its observations. The test statistic, K, will equal 0 if all groups have identical average ranks and will be positive if group ranks are different. The distribution of K when H_0 is true can be approximated by a chi-square distribution with k-1 degrees of freedom (df), where k is the number of seasons (Helsel and Hirsch, 1995). For example, with quarterly data, k = 4 and df = 3.

All N observations are given a numerical rank from 1 to N, smallest to largest. When observations are tied, the average of their ranks are assigned to each (i.e. if observations 6 and 7 have the same value, assign 6.5 as the rank for both). These ranks, R_{ij}, are then used for computation of the test statistic. Within each group, the average group rank $\overline{R}_j$ is computed as:

$$\overline{R}_j = \frac{\sum_{i=1}^{n_j} R_{ij}}{n_j}.$$

The average group rank, $\overline{R}_j$, is compared to the overall average rank, $\overline{R} = (N+1)/2$, squaring and weighting by sample size, to form the test statistic K:

$$K = \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j \left[\overline{R}_j - \frac{N+1}{2} \right]^2.$$

Reject H_0 if $K \geq \chi^2_{1-\alpha,(k-1)}$, the 1-$\alpha$ quantile of a chi-square distribution with k-1 degrees of freedom; otherwise do not reject H_0 (see Table A19 in Gilbert, 1987).

For the minimum DEQ suggested data requirements (i.e., three years of quarterly data) N=12, n_j=3, and k=4. The above equation reduces to:

$$K = \frac{1}{13} \sum_{j=1}^{4} 3 \left[\overline{R}_j - \frac{13}{2} \right]^2.$$

Table E.1 A Portion of the Quantiles of the Chi-Square Distribution with k-1 Degrees of Freedom

Degrees of Freedom (k-1)	Confidence Level	
	0.900	0.950
1	2.71	3.84
2	4.61	5.99
3	6.25	7.81
4	7.78	9.49
5	9.24	11.07
6	10.64	12.59
7	12.02	14.07
8	13.36	15.51
9	14.68	16.92
10	15.99	18.31
11	17.28	19.68

If the regulated entity discovers a seasonal trend to the collected water quality data, then this trend needs to be removed before continuing with further statistical testing. To remove the seasonal trend apply the following calculations:

1) Calculate the mean for all values from the same season in different years, $\overline{x}_k$.
2) Calculate the universal mean for all values in the data set, $\overline{X}_N$.

3) For each measurement, subtract the seasonal mean, $\bar{x}_k$ and add the universal mean, $\bar{X}_N$, to calculate the seasonally adjusted measurement. The seasonally adjusted values have lower overall variance due to removal of seasonal fluctuations.

E.2 Example

Table E.2 summarizes the Kruskal-Wallis test calculations as applied to Wells #B1 and #B2 in the example data set of Table B.1. The TDS values are ranked in ascending order from 1 to N=12. The average quarterly ranks ($R_j(1)$, $R_j(2)$, $R_j(3)$, $R_j(4)$) and the test statistic, K, are calculated, and the resulting K statistic for each well is compared to the chi-square values from Table E.1. In this case, the conclusion is that well #B1 has statistically significant seasonal variability. The above three steps for removing the seasonal variability are applied to Well #B1's data and the resulting transformation is listed in the "Adjusted B1" column in Table E.2. The result of the transformation is a data set with the same mean (254), but a significantly lower standard deviation (9.77 compared to 30.07).

Table E.2 Seasonal Testing of Example Data using Kruskal-Wallis

	Well #B1	Rank	Adjusted B1	Well #B2	Rank
Year 1 1st quarter	**305**	**12**	**266**	**252**	**7.5**
Year 1 2nd quarter	*228*	*3*	*264*	*251*	*6*
Year 1 3rd quarter	258	7	252	245	3
Year 1 4th quarter	259	8	268	252	7.5
Year 2 1st quarter	**285**	**10**	**246**	**260**	**10**
Year 2 2nd quarter	*210*	*1*	*246*	*248*	*5*
Year 2 3rd quarter	274	9	268	275	12
Year 2 4th quarter	240	5	249	272	11
Year 3 1st quarter	**290**	**11**	**251**	**256**	**9**
Year 3 2nd quarter	*216*	*2*	*252*	*246*	*4*
Year 3 3rd quarter	248	6	242	218	1
Year 3 4th quarter	235	4	244	225	2
Rj(1)		**11**			**8.8**
Rj(2)		*2*			*5*
Rj(3)		7.3			5.3
Rj(4)		5.7			6.8
K =		9.7			2.1
Critical statistic =		9.49			9.49
Seasonal variability?		Yes			No
		Adjusting for Seasonality			
Mean_1	293.3				
Mean_2	218				
Mean_3	260				
Mean_4	244.7				
Mean_total	254		254		
St_dev	30.07		9.77		

Appendix F. Secular Trends

F.1 Testing for Secular Trends Using the Mann-Kendall Test

One of the most important requirements when determining background water quality levels for the constituent(s) of concern in up-gradient monitoring wells is deciding whether temporal stationarity (steady state) exists. If the system is not in a steady state condition, then background is undefined and setting a level for comparison is not statistically valid and can lead to erroneous results.

There are several methods for determining whether the collected data show an increasing or decreasing trend through time. The U.S. Environmental Protection Agency (1988) suggests two methods. One, linear regression analysis, is somewhat simple to apply with commercial software, wherein the slope is calculated and tested for statistical significance. Though this method may be easy to perform, IDEQ does not recommend it. Linear regression is heavily influenced by outliers (Kimsey, 1996) and also makes stronger assumptions about the distribution of the data (normality of residuals, constant variance, and linearity of the relationship) (Helsel and Hirsch, 1995). Instead, IDEQ recommends that the regulated entity use the Mann-Kendall test for trend to determine if a steady state condition exists within the data.

The Mann-Kendall test is a non-parametric alternative to regression. A major advantage is that no assumption of normality is required (it is a non-parametric test). In addition, the procedure is useful if there are missing data values (e.g., a quarterly sample was missed). Data reported as less than the detection limit are assigned a common value smaller than the smallest measured value- typically the ½ the detection limit. The actual value does not matter because the test only uses the relative magnitudes of the data rather than the specific data values (Gilbert, 1987). The procedure outlined below is for cases when the number of collected background water quality data points is 40 or less (Gilbert, 1987; Gibbons, 1994). For situations where more than 40 data points are available, the regulated entity is referred to the literature (Mann, 1945; Kendall, 1975; Gilbert, 1987).

Refer to Table F.1 for the general procedure in setting up the test. First, order the data as shown by sampling date: $x_1, x_2, ..., x_N$ where x_i is the measured value for sampling date i. Second, record whether ever possible difference $x_{i'} - x_i$ is positive or negative (where i' > i), and how many total positive and negative differences occur in the data set.

Table F.1 Mann-Kendall Test Set-Up

Measurement Ordered by Time							
x_1	x_2	x_3	…	x_{N-1}	x_N	No. of + differences	No. of – differences
	$x_2 - x_1$	$x_3 - x_1$		$x_{N-1} - x_1$	$x_N - x_1$		
		$x_3 - x_2$		$x_{N-1} - x_2$	$x_N - x_2$		
				$x_{N-1} - x_3$	$x_N - x_3$		
				…	…		
				$x_{M-1} - x_{N-2}$	$x_N - x_{N-2}$		
					$x_N - x_{N-1}$		
						Total no. of + differences	Total no. of - differences

This procedure is equivalent to defining:

$$\operatorname{sgn}(x_{i'} - x_i) = \begin{cases} 1 & \text{if } x_{i'} - x_i > 0 \\ 0 & \text{if } x_{i'} - x_i = 0 \\ -1 & \text{if } x_{i'} - x_i < 0 \end{cases}$$

and computing the Mann-Kendall statistic as:

$$S = \sum_{i=1}^{n-1} \sum_{i'=k+1} \operatorname{sgn}(x_{i'} - x_i),$$

S is equal to the number of positive differences minus the number of negative differences in bottom two right-most entries of Table F.1. Conceptually speaking, if S is a large positive number, then measurements taken later in time tend to be larger than those taken earlier. Similarly, if S is a large negative number, then measurements taken later in time tend to be smaller.

Table F.2 contains part of the complete Mann-Kendall probability table (Kendall, 1975) that would be most useful for sample sizes close to the DEQ recommended minimum requirement of 12 samples. The table has been modified to perform a two-sided test (detection of either an upward or downward trend) so the probability values are twice those in the original table. The tabulated values are used to test the null hypothesis of no secular trend (statistically insignificant slope) versus the corresponding alternate hypothesis of a significant upward or downward slope. The tabulated probability corresponding to the absolute value of S is compared to the test's specified significance level (α); H_0 is rejected if the tabulated probability is less than α.

Table F.2 Values of S and corresponding probabilities for the 2-sided Mann-Kendall Test

Absolute value S	Values of N			Absolute value S	Values of N	
	12	13	16		14	15
0	1.000	1.000	1.000	1	1.000	1.000
2	0.946	0.952	0.964	3	0.914	0.922
4	0.840	0.858	0.894	5	0.830	0.846
6	0.738	0.766	0.824	7	0.748	0.770
8	0.638	0.676	0.756	9	0.688	0.698
10	0.546	0.590	0.690	11	0.590	0.626
12	0.460	0.510	0.626	13	0.518	0.548
14	0.380	0.436	0.564	15	0.450	0.496
16	0.310	0.368	0.506	17	0.388	0.436
18	0.250	0.306	0.450	19	0.330	0.380
20	0.196	0.252	0.398	21	0.280	0.328
22	0.152	0.204	0.350	23	0.234	0.282
24	0.116	0.164	0.306	25	0.192	0.240
26	**0.086**	0.128	0.266	27	0.158	0.202
28	**0.062**	**0.100**	0.228	29	0.126	0.168
30	**0.044**	**0.076**	0.194	31	**0.100**	0.140
32	**0.032**	**0.058**	0.166	33	**0.080**	0.114
34	**0.020**	**0.042**	0.140	35	**0.062**	**0.092**
36	**0.014**	**0.030**	0.116	37	**0.048**	**0.074**
38	**0.008**	**0.022**	**0.096**	39	**0.036**	**0.058**
40	**0.006**	**0.014**	**0.078**	41	**0.026**	**0.046**
42	**0.004**	**0.010**	**0.064**	43	**0.020**	**0.036**
44	**0.002**	**0.006**	**0.052**	45	**0.014**	**0.028**
46	**0.000**	**0.004**	**0.042**	47	**0.010**	**0.020**
48		**0.002**	**0.032**	49	**0.006**	**0.016**
50		**0.002**	**0.026**	51	**0.004**	**0.012**
52		**0.000**	**0.020**	53	**0.004**	**0.008**
54			**0.016**	55	**0.002**	**0.006**
56			**0.012**	57	**0.002**	**0.004**
60			**0.008**	59	**0.000**	**0.002**
62			**0.006**	61		**0.002**
64			**0.004**	63		**0.002**
66			**0.004**	65		**0.000**
68			**0.002**	67		
70			**0.002**			

Bold values indicate combinations of S and N that correspond to a trend at the $\alpha = 0.10$ level (90% confidence level).

F.2 Example

Table F.3 summarizes results for the Mann-Kendall test as applied to the example data of Table B.1. For this example, we will use a significance level of $\alpha = 0.05$. The sum of all the positive and negative tallies is made using the approach outlined in Table F.1 and an S value is computed (e.g., for Well #B1, S = 27 – 39 = -12). This S value is compared to the probability listed in Table F.2 for N = 12 and S = |-12| (p = 0.46); since 0.46 is

greater than the significance level of the test, we accept the null hypothesis (no secular temporal trend).

Table F.3 Results of Mann-Kendall test as applied to example data set

Temporal Trend						
	Well #B1	+	-	Well #B2	+	-
Year 1 1st quarter	305	0	11	252	4	6
Year 1 2nd quarter	228	8	2	251	5	5
Year 1 3rd quarter	258	4	5	245	7	2
Year 1 4th quarter	259	3	5	252	4	4
Year 2 1st quarter	285	1	6	260	2	5
Year 2 2nd quarter	210	6	0	248	3	3
Year 2 3rd quarter	274	1	4	275	0	5
Year 2 4th quarter	240	2	2	272	0	4
Year 3 1st quarter	290	0	3	256	0	3
Year 3 2nd quarter	216	2	0	246	0	2
Year 3 3rd quarter	248	0	1	218	1	0
Year 3 4th quarter	235			225		
		27	39		26	39
	S =	-12			-13	
		Table Lookup 0.46			Table Lookup 0.42	
		No trend			No trend	

Subsequent statistical analysis may proceed if the Mann-Kendall test shows that there is no statistically significant temporal trend in the background water quality data. If a significant trend exists, however, then the method in Appendix L needs to be followed in setting an interim decision threshold.

Deseasonalized data (see Appendix E.) should be used for trend analysis, because the Mann-Kendall results will be biased when seasonality is present in the data. The seasonal Kendall Test estimates the temporal trend by adjusting for seasonal variation. The test performs well when the product of the number of seasons and number of years is at least 25 (Helsel and Hirsch, 1995). For example, if three or more years of independent monthly data or seven years of quarterly data are available, the Seasonal Kendall test can be used to detect a seasonally-adjusted trend.

Appendix G. Data Pooling

G.1 Combining Well Data Sets for Normally Distributed Data

The advantage of combining background data from multiple up-gradient monitoring wells is that, by increasing sample size, greater power can be realized for decision thresholds defined at a given confidence level. However, data sets can only be combined if it can be shown that they are statistically similar.

IDEQ recommends that the data sets from two wells first be tested for similar variance using the F-test; then, if their variances are similar, they can be tested for similar means using the t-test (Larsen and Marx, 1986). For two independent, normally distributed random samples (X and Y) having means of $\bar{x}$ and $\bar{y}$ and variances of s^2_x and s^2_y, respectively, then the null hypothesis H_0 for the F-test is that $s^2_x = s^2_y$. H_0 can be rejected at the α level of significance if

$$\frac{s_y^2}{s_x^2} \text{ is either } \begin{cases} \leq F_{\alpha/2,m-1,n-1} \; or \\ \geq F_{1-\alpha/2,m-1,n-1} \end{cases}$$

where m and n are the sample size for each data set. For the case of α = 95% and m and n both equal to 12 data points (IDEQ's minimum recommended data requirement), H_0 can be rejected if s^2_y / s^2_x is either ≤ 0.29 or ≥ 3.48.

If H_0 is rejected, then the data sets cannot be combined. If H_0 cannot be rejected, then the data sets can be tested using the two-sample t-test. For two independent, normally distributed random samples (X and Y) having means of $\bar{x}$ and $\bar{y}$ and statistically equal variances of s^2_x and s^2_y, then H_0 for the t-test is $\mu_x = \mu_y$. H_0 can be rejected at the α level of significance if

$$t = \frac{\bar{x} - \bar{y}}{s_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \text{ is either } \begin{cases} \leq -t_{\alpha/2,n+m-2} \\ or \\ \geq +t_{\alpha/2,n+m-2} \end{cases}$$

where

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n+m-2} \; .$$

For the case of α = 95% and m and n both equal to 12, H_0 can be rejected if t is either ≤ -2.07 or ≥ 2.07.

If the null hypothesis is rejected, then the data sets cannot be combined.

G.2 Example

Having adjusted background well #B1 for seasonal affects and having found that both data sets (Well #B1 and Well #B2) are normally distributed (Appendix D), the question arises whether the two data sets can be combined into a single background data set. Looking at the descriptive statistics, the means are similar (254 ppm for well #B1 and 250 ppm for well #B2) and the standard deviations are similar (9.77 ppm for well #B1 and 16.4 ppm for well #B2), so the implication is that the data could be combined. To check if the variances are statistically similar, the F-test is conducted first.

$$\frac{s_2^2}{s_1^2} = \frac{16.4^2}{9.77^2} = 2.818$$

For the case of α = 95% and m and n both equal to 12 data points s^2_y / s^2_x must be $\leq$ 0.283 or $\geq$ 3.66 in order to reject H_0. In this case, H_0 cannot be rejected.

Next, the t-test is applied to the two data sets.

$$s_p^2 = \frac{(12-1)9.77^2 + (12-1)16.4^2}{12+12-2} = 182.2$$

$$t = \frac{254-250}{13.5\sqrt{\frac{1}{12}+\frac{1}{12}}} = 0.726$$

Since the test statistic (0.726) is neither $\leq$ -2.0739 nor $\geq$ 2.0739, H_0 cannot be rejected. Therefore, the TDS data sets can be combined for the purposes of setting a background water quality level.

G.3 Combining Well Data Sets for Lognormally Distributed Data

The process for determining whether background data can be combined for lognormally distributed data is the same as that for normal data, except that the calculations of the mean and standard deviation differ. The following equations can be used to calculate the arithmetic mean and standard deviation before applying the F-test and t-test (Gilbert, 1987).

$$\bar{x}_{ln} = \frac{1}{N}\sum_{i=1}^{N} \ln(x_i)$$

$$s_{ln}^2 = \frac{1}{N}\sum_{i=1}^{N} \left[\ln(x_i) - \bar{x}_{ln}\right]^2$$

Where N is the size of the sample. The decision to pool data is made using transformed data. Interpretation of the mean and standard deviation in original scale units can be obtained using the following equations:

$$\bar{x} = \exp\left(x_{ln} + s^2_{ln} \Big/ 2 \right)$$

$$s = \sqrt{\exp\left(2x_{ln} + s^2_{ln}\right)\left[\exp\left(s^2_{ln}\right) - 1\right]}.$$

G.4 Combining Well Data Sets for More Than Two Wells or Non-Parametrically Distributed Data

Levene's test (Levene, 1960) can be used to check the assumption of homogeneity of variance if there are more than two wells to be pooled. The formula is

$$W = \frac{(N-k)\sum_{i=1}^{k} N_i \left(\bar{Z}_i - \bar{Z}_{..}\right)^2}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{N_i} \left(\bar{Z}_{ij} - \bar{Z}_{i.}\right)^2}$$

where $Z_{ij} = \left|Y_{ij} - mean_i\right|$. The group means of the Z_{ij} are $\bar{Z}_{i.}$ and the overall mean of Z_{ij} is $\bar{Z}_{..}$. N is the overall sample size, k is the number of subgroups (i.e., number of background wells to be pooled), and N_i is the sample size for the i^{th} subgroup. If the data are normally distributed, all calculations are based on the original scale units. If the data are log-normal, all calculations are based on log-transformed data.

As with normal and lognormal data, non-parametric data sets need to be checked for homogeneity of variance before a comparison of medians can be performed. Levene's test can be extended (Brown and Forsythe, 1974) for working with the medians of the data sets. In the above formula, $Z_{ij} = \left|Y_{ij} - median_i\right|$ and other terms remain the same.

Levene's test rejects H_0 ($s^2_x = s^2_y$) if $W > F_{(\alpha, k-1, N-k)}$ where $F_{(\alpha, k-1, N-k)}$ is the upper critical value of the F distribution with k - 1 and N - k degrees of freedom at a significance level of α.

If H_0 is rejected, then the data sets cannot be combined. If H_0 cannot be rejected, then the Kruskal-Wallis test can be applied to determine if the subgroup medians are statistically similar and the k data sets can be combined. The Kruskal-Wallis test is described in Appendix E.

The Shapiro-Wilk Test for normality (Appendix D) can be used to determine if the combined data set is normally distributed; if it is, then the combined data can be used to set a parametric decision threshold (Appendix H or J). If the data are non-normally distributed, then the data should be logarithmically transformed and the Shapiro-Wilk test

repeated. If the data are found to be lognormally distributed, then proceed to Appendix H or J; otherwise, go to Appendix I or K.

Appendix H. Parametric Upper Tolerance Limits

H.1 The Parametric Upper Tolerance Limit as a Decision Threshold

This section of the guidance assumes that the regulated entity has a background data set that has been found to be in a steady state, to have no statistically significant seasonal effects or been corrected for seasonality, and to meet the normality or lognormality distribution assumptions. In most cases, it is assumed that the user is dealing with a new site where future water quality measurements will be compared to background water quality measurements in the same well (an intrawell analysis). The method of upper tolerance limits (UTL) is used to set the background water quality for each constituent of concern in each monitoring well.

When monitoring ground water quality, the compliance point samples are assumed to come from the same population as the background values until significant evidence of contamination can be shown (EPA, 1992a). Once the UTL is set, each compliance sample is compared to the UTL (Fisher and Potter, 1989; Gibbons, 1994; Kimsey, 1996). To minimize the false negative rate and reduce the need for verification resampling, a specified exceedance rate (coverage) is allowed (e.g., no more than 5 exceedances per 100 future comparisons). If this rate is exceeded, then significant evidence of contamination is indicated. In setting compliance limits, DEQ suggests that UTLs be set such that 95% of the tested samples (coverage) are below the limit with 95% confidence. Therefore, in this discussion, the compliance standards will be calculated for 95% confidence and 95% coverage.

Upper tolerance limits define a range within which some proportion of the population (the coverage ; in this case, 95%) will fall some proportion (in this case 95%) of the time. The limit is calculated using the following formula:

$$UTL = \bar{x} + Ks$$

where UTL = upper tolerance limit, $\bar{x}$ = arithmetic mean of the data, s = the arithmetic standard deviation of the data, and K is a constant that changes depending on the proportions used (see Gibbons, 1994 for complete mathematical formulation). Table H.1 provides examples of *K* factors for 95% coverage and 95% confidence. The process for setting tolerance intervals for lognormally distributed data is the same as that for normal data, except that the UTL is set for and compared to the log-transformed data values.

Table H.1 Partial Table of Factors (*K*) for Constructing One-Sided Normal Upper Tolerance Limits at 95% Confidence and 95% Coverage

Sample Size	95% Coverage	Sample Size	95% Coverage
8	3.188	16	2.523
9	3.032	17	2.486
10	2.911	18	2.453
11	2.815	19	2.423
12	2.736	20	2.396
13	2.670	25	2.292
14	2.614	30	2.220
15	2.566	35	2.166

See Gibbons, 1994 and Guttman, 1970 for more complete tables

H.2 Example

As was shown in Appendix G, the sample TDS data for background well #B1 and background well #B2 could be combined into a single data set of 24 data points. The resulting data set is normally distributed and has a mean of 252 ppm and a standard deviation of 23.8 ppm. The appropriate K factor, therefore, as interpolated from Table H.1, is 2.313 and the 95% UTL is:

$$UTL = 252 + 2.313 * 23.8 = 307 \text{ ppm}.$$

As long as no more than 5% of future TDS measurements exceed this threshold, the constituent of concern is deemed not to be affected by the facility's operation. Should a future exceedance of the UTL violate the 95% coverage criterion (e.g., producing 6 exceedances after 100 comparisons), then the site would be deemed out of compliance. As Gibbons (1994) points out, the UTL's specification of a coverage makes verification resampling unnecessary, because a specified number of exceedances are expected and give the method its power without a verification requirement.

If the pooled background TDS in Table B.1 happened to be lognormally distributed, then $\bar{x}$ and s of the log-transformed data in Table B.1 would be 5.53, and 0.09, respectively. The UTL for log-transformed data values would then be:

$$UTL = 5.53 + 2.313\ (0.09) = 5.74$$

Appendix I. Non-parametric Upper Tolerance Limits

I.1 The Non-parametric Upper Tolerance Limit as a Decision Threshold

For background data sets that are neither normally nor lognormally distributed, but meet the steady state condition and have been corrected for seasonal effects, a non-parametric UTL can be used to set a decision threshold for the constituents of concern. In most cases, this method will be applied to a new site where future water quality measurements will be compared to background water quality measurements in the same well (an intrawell analysis). Table I.1 shows the background sample sizes (N) required to achieve the desired coverage at varying confidence levels. The UTL is set equal to the N^{th}-highest value in the background data set.

For example, to be 85% confident (column 1, row 5) that 90% of future comparisons (column 7) will fall below the upper tolerance limit, then the 19^{th} highest background data value is used as the limit. For 95% confidence and 95% coverage, background sample size must be at least 59.

Table I.1 Sample Sizes for Non-parametric Upper Tolerance Limits

1-α	q=0.500	0.700	0.750	0.800	0.850	0.900	0.950	0.975	0.980	0.990
0.700	2	4	5	6	8	12	24	48	60	120
0.750	2	4	5	7	9	14	28	55	69	138
0.800	3	5	6	8	10	16	32	64	80	161
0.850	3	6	7	9	12	19	37	75	94	189
0.900	4	7	9	11	15	22	45	91	144	230
0.950	5	9	11	14	19	29	59	119	149	299

q = proportion of population covered by the tolerance interval. The quantity tabulated is the required sample size, N or greater, such that $P(\text{at least q of population} \le X_{(N)}) \ge 1-\alpha$.

From Conover (1999) Table A5

I.2 Example

Suppose that the TDS data in the example data set had been non-parametrically distributed. In that case, with a pooled background sample size of 24, Table I.1 would indicate that 95% coverage could be obtained at no better than a 70% confidence level (first row, eighth column). Alternatively, a 90% coverage and 90% confidence would be possible using the 22^{nd} highest TDS value in Table B.1 (285 ppm) as the non-parametric UTL.

Appendix J. Parametric Upper Prediction Limits

J.1 The Parametric Upper Prediction Limit as a Decision Threshold

This section of the guidance pertains to data sets that have been corrected for seasonality and show no statistically significant secular trend, and are normally or lognormally distributed. In addition, downgradient well water quality is assumed to differ from ambient conditions (due to current or historic land uses that have affected local site background), so that an interwell analysis is necessary (i.e., downgradient wells will be compared to up-gradient wells). In this case, a parametric upper prediction limit will be set on the basis of up-gradient water quality data (from either pooled or individual wells). As in Appendix H, the arithmetic mean and standard deviation for lognormal distributions must be correctly calculated (see Appendix G, section G.3) and the UPL defined for logarithmically transformed data.

In interwell comparison, future measurements in multiple down-gradient wells are compared to an upper prediction limit based on background water quality data from up-gradient wells (possibly pooled; see Appendix G). To reduce the sitewide false positive rate, any exceedance encountered is verified through resampling (Gibbons 1994, Sections 1.5 and 1.6). The method for verification resampling adopted by IDEQ for WLAP permits and Resource Conservation and Recovery Act (RCRA) facilities is to allow the facility to take two subsequent verification samples (these samples must also be temporally independent of the initial sample and of each other, so sufficient time must elapse between sampling and resampling to ensure temporal independence). If both samples exceed the UPL, then the initial exceedance is confirmed. This verification resampling scheme is referred to as "one of three samples in bounds" (Gibbons, 1994, Table 1.6) and has the lowest false negative rate and highest power of the three verification schemes discussed by Gibbons. Table J.1 gives multiplication factors (*K*) for the UPL formula:

$$\mathrm{UPL} = \bar{x} + Ks,$$

where $\bar{x}$ is the mean of N up-gradient background measurements and s is their standard deviation. The major difference with UPLs is that the *K* factor depends on the number of future comparisons. In general, the larger the number of future comparisons, *k*, the higher the *K* factor and the UPL; conversely, the larger the background sample size, N, the lower the *K* factor and UPL. The number of future comparisons, *k*, is defined as:

k = [number of measurements to be collected per well per year]
× [number of years] × [number of wells] × [number of COCs]

For example, for a down-gradient well that will be sampled four times a year for five years (the permit re-application period) and analyzed for two constituents of concern, *k* = 1 x 4 x 5 x 2 = 40 comparisons.

If the up-gradient, background well data cannot be pooled, then a separate UPL will be determined for each up-gradient well and decisions made as to which down-gradient wells will be compared to which up-gradient wells. These decisions will affect the value of N and *k* that are applicable for different downgradient wells because the upgradient background data sets will differ. Note that the *K* factors in Table J.1 increase only slightly for $k > 50$; if *K* factors are required for higher *k*, they can be estimated by extrapolation.

Table J.1 *K* Factors at α=0.05 for a Verification Protocol Where Both Resamples Must Confirm the Initial Exceedance

	k = Number of Future Comparisons				
N	10	20	30	40	50
4	2.02	2.42	2.65	2.82	2.94
8	1.37	1.61	1.75	1.84	1.92
12	1.21	**1.42**	1.54	**1.62**	1.68
16	1.14	1.33	1.44	1.52	1.58
20	1.10	1.28	1.39	1.46	1.51
24	1.08	**1.25**	1.35	**1.42**	1.47
36	1.03	**1.20**	1.29	**1.36**	1.41
48	1.01	**1.17**	1.27	**1.33**	1.38

From Gibbons (1994) Table 1.6, as prepared by Charles Davis based on results in Davis and McNichols (1987). Bold values indicate *K* factors that would apply to five years of quarterly future measurements (e.g., 20 comparisons of one constituent of concern for one well; 40 for two wells or for two CoCs at one well) and for various background sample sizes, each representing three years of quarterly data collected at 1, 2, 3, or 4 background wells.

J.2 Example

The pooled TDS background data set from wells #B1 and B2 has an overall N, $\bar{x}$ and s of 24, 252 and 23.8, respectively. For a 5-year permit re-application period, with quarterly samples to be collected at two down-gradient wells and a single constituent of concern (TDS), *k* is 5*4*2=40 and the K-factor read from Table J.1 is 1.42. The UPL is therefore:

$$\text{UPL} = 252 \text{ ppm} + 1.42\ (23.8) = 286 \text{ ppm}.$$

Future measurements from the down-gradient well over the monitoring period will be compared to its UPL. Every exceedance triggers a verification resampling event using the protocol specified for Table J.1. At each permit re-application, the UPL will be re-evaluated using all available data, including the new up-gradient data collected since last application as well as previous data.

If the pooled background TDS in Table B.1 happened to be lognormally distributed, then $\bar{x}$ and s of the log-transformed data in Table B.1 would be 5.53, and 0.09, respectively. The UPL for log-transformed data values would then be:

$$\text{UPL} = 5.53 + 1.42\ (0.09) = 5.66$$

Appendix K. Non-Parametric Upper Prediction Limits

K.1 The Non-parametric Upper Prediction Limit as a Decision Threshold

For data sets that are not normally or lognormally distributed, steady state, independent, and corrected for seasonal effects, a non-parametric prediction limit can be used to set background levels. Like non-parametric tolerance limits, non-parametric prediction limits require larger background sample size to provide high levels of confidence. The non-parametric UPL is set equal to the maximum value out of N independent background samples required to achieve a specified confidence level for a specified number of future comparisons. Confidence level is a function of N, the resampling plan used, and the number of future comparisons *k*. For large *k* or small α, a large number of background samples is required (Gibbons, 1994).

One assumption inherent in this procedure is that the down-gradient monitoring well's water quality data represents the same population as the up-gradient (background) well(s) to which it is compared. The interwell analysis method could be applied even at new facilities in situations where the data from down-gradient wells insufficient to justify an intrawell analysis.

As is the case with parametric UPLs, up-gradient background data from multiple wells can only be pooled if the means and standard deviations of each up-gradient well's data set are statistically indistinguishable. For non-parametric data, the statistical test to check for statistical differences of the variances and medians is Levene's Test and the Kruskal-Wallis test, respectively (Appendix G).

Table K.1 (from Gibbons, 1994, Chapter 2) summarizes confidence levels for various background sample sizes and future comparisons and the same verification resampling scheme discussed in Appendix J (take two resamples, if both also exceed the UPL then exceedance is verified). Table K.1 shows how, as background sample size increases, so does the confidence level; conversely, confidence level decreases as the number of future comparisons (*k*) increases.

Table K.1 Confidence Levels for the Non-Parametric Prediction Limit Where an Exceedance is Verified If Both of Two Resamples Also Exceed the Limit

	k = Number of Future Comparisons							
N	10	20	30	40	50	60	80	100
4	.5585	.4393	.3759	.3347	.3050	.2822	.2491	.2257
8	.7616	.6522	.5836	.5348	.4976	.4678	.4225	.3890
12	.8538	.7676	.7072	.6613	.6246	.5942	.5463	.5095
16	.9023	.8356	.7852	.7449	.7115	.6831	.6368	.6001
20	.9305	.8785	.8369	.8024	.7729	.7473	.7044	.6695
25	.9516	.9126	.8798	.8516	.8268	.8047	.7668	.7350
35	.9729	.9492	.9279	.9087	.8912	.8751	.8463	.8211
50	.9858	.9727	.9604	.9488	.9379	.9275	.9083	.8908

From Gibbons (1994) Table 2.13, as prepared by Charles Davis based on results in Davis and McNichols (1993)

K.2 Example

The pooled sample size for background TDS data from upgradient wells #B1 and B2 is N = 24; for a single constituent of concern and a 5-year permit re-application period with quarterly samples to be collected at two down-gradient wells, *k* is 40. The UPL is set equal to the highest value of the N background measurements (305 ppm, Table B.1). Table K.1 shows that any future comparisons to this UPL would be at no higher than about an 85% confidence level. If a higher confidence level were desired, then fewer future comparisons would have to be specified; for example, to achieve a 90% confidence with N = 24, *k* would have to be limited to about 24 comparisons (interpolating the sixth row of Table K.1). This would correspond to a 3-year comparison period in the above example.

Appendix L. Interim Decision Thresholds in the Presence of a Secular Trend

L.1 Introduction

The purpose of this appendix is to provide guidance for setting a regulatory threshold in cases where (1) background water quality cannot be established because the up-gradient monitoring wells exhibit secular trends in water quality, and/or (2) site practices are being modified to bring the system into compliance and down-gradient water quality is or will be affected.

In case (1) the facility is affected by secular trends originating off site and over which the regulated entity may have no control. In case (2), previous permitted use of a facility may have caused down-gradient wells to exceed the primary and/or secondary constituent standards (IDAPA 58.01.11.200) or ACLs for constituent(s) of concern, and the regulated entity may be implementing operational changes in response. Future water quality trends may develop as site background comes to a new steady state condition. In either case, the regulated entity should continue monitoring the site during the transition period until methodologies in Appendices H - K can once again be applied. In the interim, the method described in Section L.2 should be followed.

L.2 Procedure for setting a decision threshold under non-steady state conditions

The assumption is made that if ground water is not currently in a steady state condition, then it is approaching a steady state condition because upgradient land uses or practices at the facility have stabilized. During this transition time, data still need to be collected, and limits are required to ensure that the approved practices are causing the water quality to continue to trend towards a future steady state condition. The following method is to be used for setting compliance limits during the transition time.

1) Each year, in the first quarter, determine if the system is trending or in steady state (Mann-Kendall Test).
   - The facility should recheck annually because it is likely that a trend will change with time as a new steady state condition is approached. For example, the constituent(s) of concern will begin to level off as they approach steady state.
   - Once the COC has defined a statistically steady state condition, use the last 12 data points to establish tolerance or prediction intervals described in the appendices.
   - In cases where there is a new, mutually agreed-upon mean concentration for a particular constituent of concern (case-by-case basis), the standard deviation should be based on the background concentration in the monitoring wells.
2) If the system shows a temporal trend, estimate the trend by Sen's slope method outlined in Section L.3.
3) In cases where a limit is needed for comparison with the trending data, use the 100(1-α)% lower confidence limit for an increasing trend and the 100(1-α)% upper

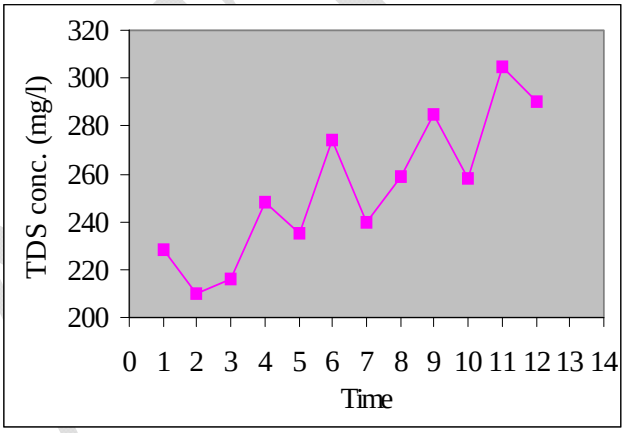
confidence limit for a decreasing trend. The difference between the next measurement and its previous measurement should be within these limits.

4) In the case of an increasing trend, the regulated entity should use any exceedance as a warning that the current practices being used to reach a new steady state condition may not be adequate. An exceedance of the confidence limit during the transition period will be addressed on a case-by-case basis.

L.3 Non-parametric Sen's slope method to estimate trend

Sen's trend estimator is simple and particularly useful for groundwater monitoring (Gibbons, 1994). The following fabricated example shows how to calculate Sen's slope for a trending groundwater TDS time series. The data represent a modified version of well B1's data from a previous example.

Time ID	Fabricated Well B1
Year 1 1st quarter	228
Year 1 2nd quarter	210
Year 1 3rd quarter	216
Year 1 4th quarter	248
Year 2 1st quarter	235
Year 2 2nd quarter	274
Year 2 3rd quarter	240
Year 2 4th quarter	259
Year 3 1st quarter	285
Year 3 2nd quarter	258
Year 3 3rd quarter	305
Year 3 4th quarter	290



Step 1: Lay out the concentrations in temporal order

Step 2: Obtain individual slopes Q_i for each period i' > i, where i is any starting time and i' is a time following i. The total number of individual slope comparisons is N'=n(n-1)/2. In this example, N'=(12)(11)/2=66.

Time period	1	2	3	4	5	6	7	8	9	10	11	12
TDS conc. (mg/l)	228	210	216	248	235	274	240	259	285	258	305	290
Q_i		-18.0	-6.00	6.67	1.75	9.20	2.00	4.43	7.13	3.33	7.70	5.64
			6.00	19.0	8.33	16.0	6.00	8.17	10.7	6.00	10.6	8.00
				32.0	9.50	19.3	6.00	8.60	11.5	6.00	11.1	8.22
					-13.0	13.0	-2.67	2.75	7.40	1.67	8.14	5.25
						39.0	2.50	8.00	12.50	4.60	11.7	7.86
							-34.00	-7.50	3.67	-4.00	6.20	2.67
								19.0	22.5	6.00	16.3	10.0
									26.0	-0.50	15.3	7.75
										-27.0	10.0	1.67
											47.0	16.0
												-15.0

Step 3: Rank the N' individual slopes from smallest to largest. In this example, the ranking results are shown in the following table:

Q_i	Rank	Q_i	Rank	Q_i	Rank	Q_i	Rank
-34.0	1	3.33	18	7.75	35	11.7	52
-27.0	2	3.67	19	7.86	36	12.5	53
-18.0	3	4.43	20	8.00	37	13.0	54
-15.0	4	4.60	21	8.00	37	15.3	55
-13.0	5	5.25	22	8.14	39	16.0	56
-7.50	6	5.64	23	8.17	40	16.0	56
-6.00	7	6.00	24	8.22	41	16.2	58
-4.00	8	6.00	24	8.33	42	19.0	59
-2.67	9	6.00	24	8.60	43	19.0	59
-0.50	10	6.00	24	9.20	44	19.3	61
1.67	11	6.00	24	9.50	45	22.5	62
1.67	11	6.00	24	10.0	46	26.0	63
1.75	13	6.20	30	10.0	46	32.0	64
2.00	14	6.67	31	10.6	48	39.0	65
2.50	15	7.13	32	10.7	49	47.0	66
2.67	16	7.40	33	11.1	50		
2.75	17	7.70	34	11.5	51		

Step 4: The trend estimate, *S*, is the median of the individual slopes. If N' is odd, S is the middle slope, $Q_{(N'+1)/2}$; if N' is even, $S=1/2*(Q_{(N'/2)}+Q_{(N"+2)/2})$. In this example, the estimated trend slope is the average of the 33rd and the 34th Q_i (highlighted in yellow). Therefore $S= (Q_{33}+Q_{34})/2=(7.4+7.7)/2=7.55$ mg/l. Therefore, the TDS concentration was increasing for the three year monitoring period, and the estimated rate of increase is 7.55 mg/l per quarter.

Step 5: calculate the variance of the estimated slope using the following formula (Kendall, 1975):

$$\text{var}(S) = \frac{1}{18}[n(n-1)(2n+5) - \sum_{p=1}^{q} t_p(t_p-1)(2t_p+5)]$$

where n is the sample size, q is the number of values that have ties and t_p is the number of tied measurements (highlighted in gray) for a particular value. In this example, n=12, q=6 and t_p are 2,6,2,2,2,2 for each tied value. Therefore,

$$\sum_{p=1}^{q} t_p(t_p-1)(2t_p+5)] = (2*1*9)+(6*5*17)+(2*1*9)*4=600$$

Thus,

$$\text{var}(S) = \frac{1}{18}[(12)(11)(29)\text{-}600]=179.3$$

Step 6: Calculate the lower confidence limit (L.C.L.) for an increasing trend or the upper confidence limit (U.C.L.) for a decreasing trend.

$$L.C.L:$$

$$M_1 = \frac{N' - Z_{1-\alpha}\sqrt{\mathrm{var}(S)}}{2}$$

$$U.C.L:$$

$$M_2 = \frac{N' + Z_{1-\alpha}\sqrt{\mathrm{var}(S)}}{2}$$

Z is the score of a standard normal population with mean = 0, and standard deviation = 1. M_1 and M_2 are the orders for the ranked individual slopes in Step 3. If M_1 and M_2 are not integers, interpolation can be made from the neighboring two ranked slopes. For example, if M_1=3.7, the neighboring two ranks are 3 and 4. Weighted average of the individual slopes from rank 3 and rank 4 are used to estimate the slope at rank 3.7. As 3.7 is closer to 4, individual slope at rank 4 gives more weight. Therefore, the L.C.L = (4 - 3.7)*Q_3+(3.7 - 3)*Q_4. In this example, *S* is positive, indicating an increasing trend. Therefore the 95% L.C.L is:

$$M_1 = \frac{66 - 1.65\sqrt{179.3}}{2} = 21.95$$

and $Q_{21.95}$=0.05*Q_{21}+0.95Q_{22} = 5.22 mg/l

Therefore, the next measurement should be no greater than 5.22 mg/l higher than its immediately prior measurement.

Appendix M: Example Scenario for an Existing WLAP Facility with No Chemical Impact

The majority of the WLAP facilities using this guidance will probably be in this category. In addition to existing facilities, new facilities where previous land uses have altered the down-gradient ground water at compliance wells from an ambient condition (Figure 2.2) also fall into this category. As with other facilities, the first steps are to conduct descriptive statistics on the constituent(s) of concern for each background well (Chapter 3).

Following the initial descriptive statistical documentation (Appendix B) and an evaluation of data independence (Appendix C), the distribution of the each constituent of concern should be checked for normality / lognormality (Appendix D) and then its temporal behavior evaluated for statistically significant secular trends and seasonal pattern (Appendix E). The concentration versus time diagrams will likely indicate whether there is a cyclic nature to the data, but seasonality must be statistically demonstrated. Ideally, at least three years of quarterly data should be available for this analysis (wherein each quarter is tested in the same month). Some of the variation may be due to changing land uses (nearby agricultural, river flow, canal flow, etc.) as well as true seasonal effects such as precipitation patterns, evapotranspiration, etc. The preferred method for determining seasonal stationarity is the non-parametric Kruskal-Wallis test (Appendix E). Once seasonality has been tested for and possibly removed, the resulting data sets should be tested for secular trends using the recommended non-parametric Mann-Kendall test (Appendix F).

If the Mann-Kendall test shows no temporal trend for background water quality data, then the methodology in Appendix G should be used to determine whether data from multiple background wells can be pooled. If the Mann-Kendall test shows that there is a temporal trend, then an alternative method needs to be followed to define a decision threshold for future monitoring (Appendix L).

After defining background water quality for each constituent of concern, decision thresholds for future monitoring are set. In most cases, the existing facility will have altered background water quality, and the process outlined in Appendix J can be used to set parametric prediction levels for future interwell comparisons for the constituents of concern. To do so, the data set must (1) exhibit no temporal trends, (2) have no statistically significant seasonal effects or be corrected for seasonality, and (3) be normally or lognormally distributed. Appendix K makes the same assumptions except that the data distribution is non-parametric. In either case, site conditions are such that down-gradient water quality has been affected by the facility, requiring that interwell statistical methods be applied.

Wherever possible, site conditions should be evaluated to determine if interwell comparisons are justified. For example, in situations where background water quality is highly variable, or aquifer heterogeneity makes it difficult or impossible to decide which

upgradient well(s) should be compared to a downgradient well, then intrawell comparison procedures or modifications to those suggested in this document should be considered. For example, can background data in downgradient wells be filtered of outliers (Appendix N.4) that may represent existing site impacts, prior to applying intrawell comparisons? Or could alternative methods for setting decision thresholds be used, such as Shewart-CUSUM control charts (Gibbons, 1994)? Such a decision may prove to be far more defensible for an existing facility than trying to force interwell comparisons where hydrogeologic conditions do not warrant them.

The single greatest advantage to using intrawell methods (including variants such as Shewart-CUSUM charts) is that compliance decisions are solely based on the statistical behavior of constituents of concern in individual wells rather than between wells whose up- vs. downgradient hydrogeologic relationship may be suspect or unknown. In all cases, the use of intrawell comparisons at existing facilities are justified only if the regulated entity can demonstrate that the data set(s) to be considered as "background" for the constituents of concern in down-gradient wells have either not been affected by the facility's prior operations or appropriately filtered of suspected contamination influences (e.g., outliers).

Appendix N: Applying Intrawell Analysis at Existing Facilities When Interwell Methods are Inadvisable

At existing facilities, large variations in natural background water quality across a site can make it difficult to identify hydrologically appropriate pairs of wells for interwell comparison. This problem that can be exacerbated by very slow ground water flow rates between wells. This increases the difficulty of identifying true exceedances from other confounding influences in an interwell comparison. For these reasons intrawell comparison, where it can be justified, is the method of choice for compliance monitoring (Gibbons, 1994).

At existing facilities intrawell comparison is preferred over interwell analysis whenever possible. Specifically, if historical background data in a downgradient well demonstrates that the constituents of concern have not been impacted by the facility's operations, then the use of intrawell methods may be justified for detection monitoring.

To apply an intrawell monitoring method at an existing facility, the regulated entity must demonstrate that preexisting contamination was not present in the downgradient wells during the historical period selected for establishing the background level. The intrawell method DEQ suggests for monitoring preexisting facilities is the combined Shewhart-CUSUM control chart method. It is capable of detecting both immediate and gradual releases and is applicable to data sets containing up to 75% non-detects. It combines the power of the Shewhart control chart method, which is ideal for rapid detection of large releases, and the Cumulative SUM method that is sensitive to gradual releases. Data must be temporally independent, so that quarterly data are recommended, and should be screened for outliers or other evidence of preexisting impacts by the facility.

N.1 Demonstrating that intrawell comparison is appropriate for site-specific conditions

The regulated entity should provide evidence that COCs in downgradient wells have not been affected by the facility. For example, based on an evaluation of historical data (outlier screening, seasonality, trends), the regulated entity may be able to demonstrate that a window of time exists for defining background COC levels for each well in the monitoring network. Outlier detection is addressed in Section N.4. To check trend and seasonality of the historical data, refer to Appendix E and F.

Alternatively, groups of up- and downgradient wells can be tested for statistical similarity (using methods of Appendix G) to identify those downgradient wells whose water quality is statistically indistinguishable from up-gradient wells and whose future data could be analyzed using either interwell or intrawell methods.

For each downgradient well that has not been affected by facility impacts, screen the historical COC data and remove outliers to establish an historic statistical baseline (e.g.: Gibbons, 1994, section 8.4.3, p.164-165). Justify using an intrawell comparison by

demonstrating that no COCs have yet been detected in the downgradient well and that other indicator constituents show no significant trends (ASTM, 1998; Cal EPA, 2001).

N.2 Apply a Shewhart-CUSUM control chart method to detect future changes in water quality

The Shewhart-CUSUM control chart procedure is a widely used intrawell comparison method that EPA recommends for identifying a statistically significant increase in chemical concentrations at a single monitoring location (EPA, 1989, 1992; ASTM, 1998; ITRC, 2006; URS, 2003; Gibbons, 1994, 1999). DEQ recommends twelve background samples are needed to compute a standardized difference value and control limits against which subsequent measurements from the same well are compared. The method has been applied in an evaluation mode at various sites (e.g., Chou, 2004) and has performed well. Because the method is sensitive to both gradual (long-term) and sudden (short-term) increases, it allows for detection of facility impacts at different spatial and temporal scales. The method is applicable to data that are independent and normally distributed; hence a well's historic background data should be evaluated for temporal independence, or the analysis should be restricted to data that have been collected no more frequently than quarterly.

The procedure can be implemented as follows: Let x_i be a series of independent background observations i = 1, 2…., n (n = 12 at minimum). Let x_j be a series of future monitoring measurements j = 1, 2, 3….. . Then, using the background data, the following steps are applied:

1. Check the data for normality and temporal independence, and apply an appropriate transformation if necessary. For transformed data, the following steps will be performed on transformed data.
2. Use the background data (x_i) to compute $\bar{x}$ and s as estimates for the mean μ and standard deviation σ of the normal distribution.
3. Define three parameters (all in units of standard deviation) for the control chart:
   SCL – Shewhart Control limit
   h – CUSUM Control limit
   k – amount of shift in the mean to be detected rapidly

   For ground water quality monitoring, experience has shown that a combination of parameter values for SCL = 4.5, h = 5.0 and k = 1 are most appropriate. Other values may also be used, depending on the sampling scheme and the sample size (Gibbons, 1994).

4. For each future data value, compute its standard normal deviate, z_j:

$$z_j = \frac{x_j - \bar{x}}{s}$$

5. Compute the CUSUM statistic S_j

$$S_j = \max[0, (z_j - k) + S_{j-1}] \; ; \; S_0 = 0.$$

If either x_j > SCL **or** S*j* > h, then verification resampling is conducted (these samples must be temporally independent of the initial sample, so sufficient time must elapse between sampling and resampling to ensure temporal independence). This timeframe should be determined based on consideration of site-specific ground water flow conditions and after consultation with DEQ. . A well is determined to be out of compliance only if the verification result also exceeds either the SCL **or** h. If verification resampling is implemented during monitoring, its analytical result is used in formulas (4) and (5) to update the CUSUM statistic for future comparisons.

Note that the Shewhart portion of the test quickly detects large, rapid deviations from background, whereas the CUSUM portion of the combined test is sequential; a small positive shift in the mean concentration over the preceding time period will slowly aggregate in the CUSUM statistic and eventually cause the test statistic to exceed the CUSUM control limit h.
Thus, the combined Shewhart-CUSUM method has the ability to detect rapid as well as gradual releases from a monitored facility.

N.3 Example

For this example, we assume that the fabricated data in the table have been screened and outliers removed, corrected for seasonality and are free of any secular trend.

Table N-1. Background TDS measurements

Background sample n	Background TDS, mg/L
1	259
2	228
3	240
4	216
5	285
6	235
7	290
8	274
9	290
10	228
11	216
12	248

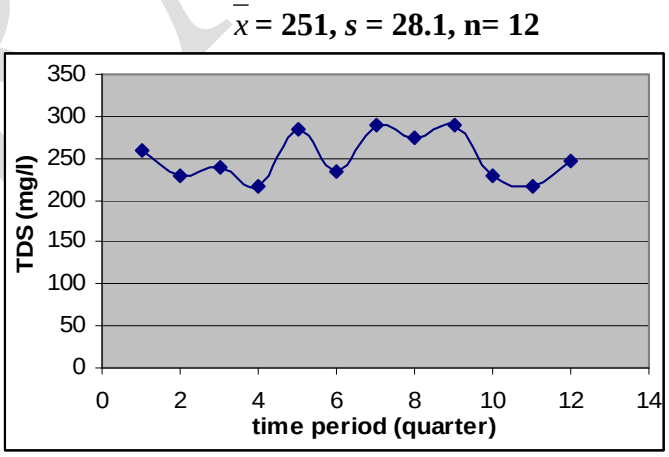


Table N-2. Monitoring TDS measurements

Monitoring sample	Measured TDS (mg/l)	z_i	z_i-k	S_i
Year 1 1st quarter	258	0.26	-0.74	0.0
Year 1 2nd quarter	305	1.93	0.93	0.9
Year 1 3rd quarter	289	1.36	0.36	1.3
Year 1 4th quarter	268	0.61	-0.39	0.9

Set h= 5, SCL= 4.5, k= 1, calculate z_i, z_i-k and S_i as outlined in section N.2. The values are summarized in Table 2. Shewhart-CUSUM control chart is shown in the Figure following. Figure N-1 shows that both z_i and S_i are within specified limits. Therefore, this one-year monitoring shows the system is in compliance.

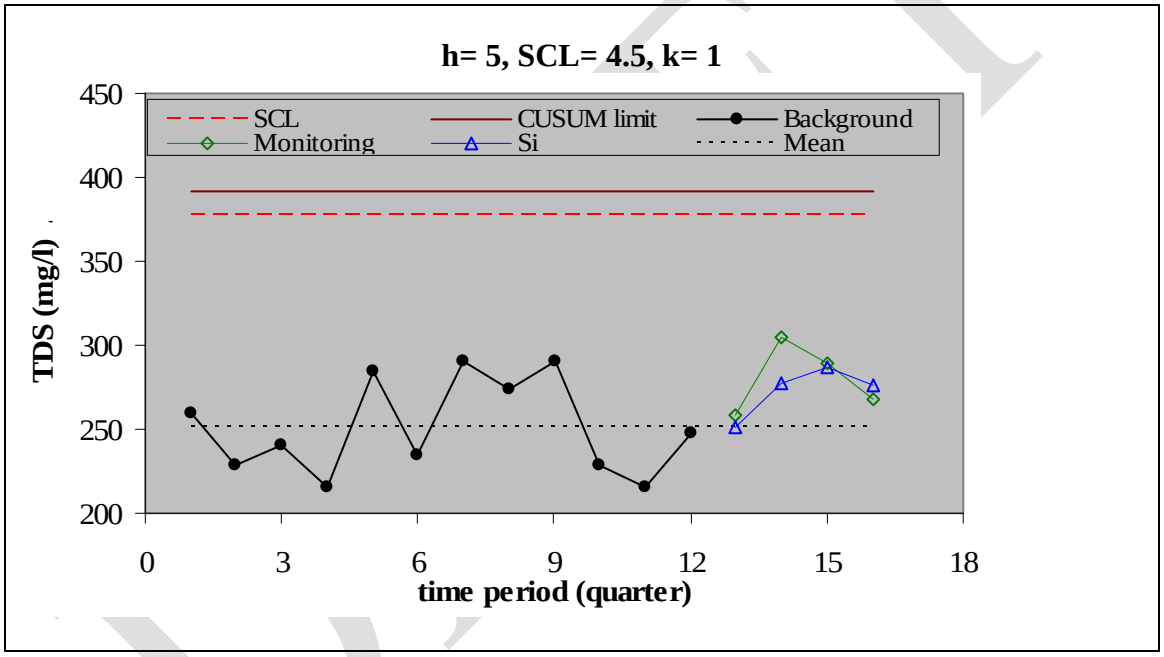


Figure N.1 Comparison to historical data and specified limits

N.4 Detection of Outliers in Background Data

The following discussion outlines the steps necessary to detect outliers using Dixon's method. Dixon's test can be used when the number of suspected outliers is small. If m outliers are suspected, all m tests must be performed regardless of the outcomes of the previous m-1 test. If the m^{th} test exceeds the critical value, all m outliers must be rejected. If data are not normal in original scale, proper transformation should be applied. Once the data are transformed, the following steps then should be applied.

1. Sort the data from lowest to highest, denoted by $x_{(i)}$ where i=1 to n.
2. Calculated the average of the data, $\bar{x}$

3. Calculated $\left|x_{(i)} - \bar{x}\right|$ for each observation and sort the difference from largest to smallest.
4. Decide the number of suspected outliers, m
5. Calculate Dixon's statistics using following formula (Gibbons, 1994) for the m outliers, starting from the most extreme value.

n	Highest value	Lowest value
3-7	$\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(1)}}$	$\frac{x_{(2)} - x_{(1)}}{x_{(n)} - x_{(1)}}$
8-10	$\frac{x_{(n)} - x_{(n-1)}}{x_{(n)} - x_{(2)}}$	$\frac{x_{(2)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$
11-13	$\frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(2)}}$	$\frac{x_{(3)} - x_{(1)}}{x_{(n-1)} - x_{(1)}}$
14-25	$\frac{x_{(n)} - x_{(n-2)}}{x_{(n)} - x_{(3)}}$	$\frac{x_{(3)} - x_{(1)}}{x_{(n-2)} - x_{(1)}}$

6. Compare the statistic to following tabulated critical values (Gibbons, 1994) and draw conclusions.

n	5% level	1% level	n	5%level	1%level
3	.941	.988	14	.546	.641
4	.765	.889	15	.525	.616
5	.642	.780	16	.507	.595
6	.560	.698	17	.490	.577
7	.507	.637	18	.475	.561
8	.554	.683	19	.462	.547
9	.512	.635	20	.450	.535
10	.477	.597	21	.440	.524
11	.576	.679	23	.421	.505
12	.546	.642	24	.413	.497
13	.521	.615	25	.406	.489

Using the same fabricated data, we assume that there is one more observation in the historical data, the 13th measurement with TDS equals 380 mg/l. Applying above outlined steps results in Table 3. The ascending sorted observations x(i) is shown in column 3. $\bar{x}$ equals 260.7 mg/l. Therefore sorted $\left|x - \bar{x}\right|$ for each observation and its corresponding measured values are shown in column 4 and 5. Suspected number of outlier is 3 (m=3), the highest TDS and the two lowest TDS in the data set. Using Dixon's formula in Step 5 for n=12, starting with the most extreme value TDS=380 mg/l, Dixon's statistic is (380-290)/(380-216)=0.549. It is significant at 5% level but not at 1%

level comparing to critical values in the Table of Step 6. Continuing with the lowest TDS=216, using the same approach, Dixon's statistic is 0.162, not significant at 5% level and 1% level. Therefore, the observation with TDS=380 mg/l can be rejected and observations with TDS=216 should be retained for intrawell comparison.

Table N-3. Background TDS measurement with fabricated outlier

Background Sample, n	Background TDS, mg/l	Sorted TDS $x_{(i)}$		Sorted $\lvert x - \bar{x} \rvert$	Corresponding Background TDS, mg/l	Dixon's Statistic
1	259	$x_{(8)}$		119.3077	**380**	0.549*
2	228	$x_{(3)}$		44.69231	216	0.162
3	240	$x_{(6)}$		44.69231	216	0.162
4	216	$x_{(1)}$		32.69231	228	
5	285	$x_{(10)}$		32.69231	228	
6	235	$x_{(5)}$		29.30769	290	
7	290	$x_{(11)}$		29.30769	290	
8	274	$x_{(9)}$		25.69231	235	
9	290	$x_{(12)}$		24.30769	285	
10	228	$x_{(4)}$		20.69231	240	
11	216	$x_{(2)}$		13.30769	274	
12	248	$x_{(7)}$		12.69231	248	
13	**380**	$x_{(13)}$		1.692308	259	

References

Aitchison, J., 1955; "On the distribution of a positive random variable having a discrete probability mass at the origin"; J. Amer. Statist. Assoc.; Vol. 50, pp. 901-908.

American Society for Testing and Materials (ASTM), 1998, Standard Guide for Developing Appropriate Statistical Approaches for Groundwater Detection Monitoring Programs; Designation: D 6312-98, pp. 1325-1338. West Conshohocken, Pennsylvania: ASTM.

Barcelona, M. J., H. A. Wehrman, M. R. Schock, M. E. Sievers, J. R. Karny. Sampling Frequency for Ground-Water Quality Monitoring. Report #EPA/600/4-89/032, Washington, D.C., 1989.

Bertolino, F., Luciano, A., and Racugno, W., 1983, Some aspects of detection networks optimization with the kriging procedure; Metron, 41(3): 91-107.

Brown, M.B. and Forsythe, A.B. 1974. Journal of the American Statistical Association. Volume 69. Pages 364-367.

Cameron, K, and Hunter, P., 2002, Using spatial models and kriging techniques to optimize longterm ground-water monitoring networks: a case study, Environmetrics, Vol 13, 629-656.

Chou, C. J., 2004, Evaluation of an alternative statistical method for analysis of RCRA groundwater monitoring data at the Hanford site; PNNL-14521, Pacific Northwest National Laboratory; Richland, Washington.

Chou, C.J., O'Brien, R.F. and Barnett, D.B., 2001, Application of Intrawell Testing of RCRA Groundwater Monitoring Data When No Upgradient Well Exists; J. Environmental Monitoring and Assessment, Vol. 71, Number 1, pp. 91-106.

COHEN, A.C., 1991; "Truncated and censored samples: Thoery and Applications"; Marcel Dekker, NY. 312 pp.

Conover, W.L. 1999. Practical Nonparametric Statistics, 3rd edition. John Wiley & Sons, New York, NY.

Crow, E.L., Davis, F.A., and Maxfield, M.W. 1960. Statistics Manual. Dover Publications Inc., New York, NY.

Cressie, N.A.C. 1993. Statistics for Spatial Data. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, NY.

Davis, C.B. and McNichols, R.J. 1987. One-sided Intervals for at Least *p* of *m* Observations from a Normal Population on Each of *r* future occasions. Technometrics. Volume 29. Pages 359-370.

Davis, C.B. and McNichols, R.J. 1993. Nonparametric Simultaneous Prediction Limits. Technical Report, Environmetrics and Statistics LTD, Henderson, NV.

Domenico, P.A. and Schwartz, F.W. 1990. Physical and Chemical Hydrogeology. John Wiley & Sons, Inc. New York, NY.

EPA (U.S. Environmental Protection Agency). 1988. Statistical Methods for Evaluating the Attainment of Superfund Cleanup Standards, Volume 2: Groundwater. Draft 2.0. Prepared by Westat Inc. under contract No. 68-01-7359.

EPA (U.S. Environmental Protection Agency). 1992a. Statistical Analysis of Groundwater Monitoring Data at RCRA Facilities: Addendum to Interim Final Guidance. EPA/530-R-93-003.

EPA (U.S. Environmental Protection Agency). 1992b. Methods for Evaluating the Attainment of Cleanup Standards, Volume 2: Groundwater. Prepared by Environmental Statistics and Information Division. EPA/230-R-92-014.

Fetter, C.W. 1993. Contaminant Hydrogeology. Macmillan Publishing Company. New York, NY.

Fisher, S.R. and Potter, K.W. 1989. Methods for Determining Compliance with Groundwater Quality Regulations at Waste Disposal Facilities. Submitted to the Wisconsin Department of Natural Resources.

Gibbons, R.D. 1994. Statistical Methods for Groundwater Monitoring. John Wiley & Sons, New York, NY.

Gibbons, R.D., 1999, Use of Combined Shewart-CUSUM Control Charts for Ground Water Monitoring Applications: Ground Water, Vol. 37, No. 5, pp. 682-691.

Gilbert, R.O. 1987. Statistical Methods for Environmental Pollution Monitoring. Van Nostrand Reinhold, New York, NY.

Guttman, I. 1970. Statistical Tolerance Regions: Classical and Bayesian. Hafner Publishing, Darien, Connecticut.

Harris, J., Loftis, J.C., and Montgomery, R.H. 1987. Statistical Methods for Characterizing Ground-Water Quality. Ground Water. Volume 25, No. 2. Pages 185-193.

Helsel, D.R., 1990 “Less than obvious: Statistical treatment of data below the detection limit”, Environmental Science and Technology.

Helsel, D.R. and Hirsch, R.M. 1995. Statistical Methods in Water Resources. Studies in Environmental Science 49. Elsevier, New York, NY.

Helsel, D.R., 2005. Nondetects and Data Analysis: Statistics for Censored Environmental Data. John Wiley & Sons, Inc. publishing.

Interstate Technology & Regulatory Council (ITRC), 2006, Evaluating, Optimizing, or Ending Post-Closure Care at MSW Landfills Based on Site-Specific Data Evaluations; Washington, D.C.: Interstate Technology & Regulatory Council, Alternative Landfill Technologies Team, www.itrcweb.org.

Isaaks, E.H. and Srivastava, R.M., 1989, Applied Geostatistics; Oxford University Press, New York.

Johnson, V.M., Tuckfield, R.C., Ridley, M.N., & Anderson, R.A., 1996, Reducing the sampling frequency of ground-water monitoring wells; Environmental Science & Technology, 30(1): 355-358.

Kendall, M.G. 1975. Rank Correlation Methods, 4th edition. Charles Griffon, London.Kimsey, M.B. 1996. Implementation Guidance for the Ground Water Quality Standards. Prepared for the Washington State Department of Ecology Water Quality Program, Olympia, Washington.

Larsen, R.J. and Marx, M.L. 1986. An Introduction to Mathematical Statistics and Its Applications: 2nd edition. Prentice-Hall, New Jersey.

Lehmann, E.L. 1998. Nonparametrics, Statistical Methods Based on Ranks. Holden-Day, Oakland, California.

Levene, H. 1960. *In* Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, I. Olkin et al. eds., Stanford University Press, Stanford, California. Pages 278-292.

Manly, B.F.J. and Mackenzie, D.I., 2003, CUSUM environmental monitoring in time and space; Environmental and ecological Statistics, Vol. 10, pp. 231-247.

Mann, H.B. 1945. Nonparametric Tests Against Trend. Econometrica. Volume 13. Pages 245-259.

Moore, D. S. and G. P. McCabe, 1998. Introduction to the practice of statistics. W.H. Freeman and Company

Ogden, A.E. 1987. A Guide to Groundwater Monitoring and Sampling. Idaho Department of Health and Welfare: Division of Environment. Water Quality Report No. 69, Boise, Idaho.

Ohio EPA, 2006, Hydrogeologic Site Investigation Questions, OAC RULE 3745-27-06 (C)(3);
http://www.epa.state.oh.us/dsiwm/document/newsPDFs/gw_rule_changes_faqs.pdf

Oswina, A., U. Lall, T. Sangoyomi, K. Bosworth, 1992, Methods for Assessing the Space and Time Variability of Groundwater Data; U.S. Geological Survey PB-94 116548, Washington, D.C.

Ott, R. L. and W. Mendenhall, 1995 Understanding Statistics. International Thomson Publishing

Ridley, M. and D. MacQueen, 2005, Cost-Effective Sampling of Groundwater Monitoring Wells: A Data Review & Well Frequency Evaluation; Lawrence Livermore National Laboratories, UCRL-CONF-209770.

Royston J.P.; 1993, "A Toolkit for Testing for Non-Normality in Complete and Censored Samples"; The Statistician; Vol. 42, No. 1. pp. 37-43.

Ryan, T. A. and Joiner, B. L., 1976, Normal Probability Plots and Tests for Normality, The Pennsylvania State University, http://www.minitab.com/resources/articles/normprob.aspx, website accessed July 7, 2007

Shapiro, S.S. and Wilk, M.B. 1965. "An Analysis of Variance Test for Normality (complete samples)." *Biometrika*. Volume 52. Pages 591-611.

SPSS Inc. 2000. SYSTAT 10. Chicago, Illinois.

U.S. Environmental Protection Agency, 1989, Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities, Interim Final Guidance; EPA/530 - SW-89-026, Washington, D. C.

U.S. Environmental Protection Agency, 1992, Statistical Analysis of Ground-Water Monitoring Data at RCRA Facilities. Addendum to Interim FinalGuidance; EPA/530-R-93-003, Washington, D. C.

URS Group, 2003, Implementation of Alternative Measures Industrial Waste Lagoon, Tooele Army Depot, Tooele, UT; Final System Non-operation Test Proposal; URS Group, Inc., Bethesda.

Virginia Department of Environmental Quality. 2003. Data Analysis Guidelines for Solid Waste Facilities

Washington State, Department of Ecology, 2005, Implementation Guidance for the Ground Water Quality Standards; Olympia, Washington.

Acronym/Symbol Definition List

α	False rejection (or false positive) decision error
ACL	Alternative concentration limit
b_1	Slope of the linear regression line
b_0	Intercept of the linear regression line
COC	Constituents of concern
CV	Coefficient of variation
γ	Skewness
IDEQ	Idaho Department of Environmental Quality
EPA	U.S. Environmental Protection Agency
F	Variance ratio from the table of the F-distribution
H_0	Null hypothesis
H_A	Alternative hypothesis
IQR	Interquartile Range
K	Kruskal-Wallis (K-W) test statistic
k	Number of seasons (typically 4 for the K-W seasonality test)
K	Multiplier used for setting UTLs or PLs
k	The number of future comparisons
m	The number of years for which data were collected
MSE	Mean square error
N	The sample size or total number of measurements (= n x m)
n	The number of measurements per year (quarterly = 4)
ppm	Parts per million
r^2	Coefficient of determination
$\overline{R}_j$	Average group rank for Kruskal-Wallis test
s	The standard deviation of a sample data set
S	The Mann-Kendall test statistic
s^2	The variance of a sample data set
SSE	Sum of squares due to error
$s_{\bar{x}}$	Standard error
TDS	Total dissolved solids
UPL	Upper prediction limit
UTL	Upper tolerance limit
W	Shapiro-Wilk test statistic
W	Levene test statistic
WLAP	Wastewater land application permit
x_i, y_i	Constituent concentration for the i^{th} ground water sample
$\overline{x}$ or $\overline{X}_N$	The mean (or average) of a sample data set
$\overline{x}_k$	The mean for all values from the same month but different years
x_{jk}	An alternative way of denoting a chemical measurement, where k = 1, 2, …, m denotes the year, and j = 1, 2, …, n denotes the sampling period (season) within the year. The subscript for x_{jk} is related to the subscript for x_i in the following manner: i = (k-1)n + j.
$\chi^2_{1-\alpha,(k-1)}$	The 1-α quantile of a chi-square distribution with k-1 degrees of freedom